

(19)日本国特許庁(JP)

(12)公表特許公報(A)

(11)公表番号
特表2024-514691
(P2024-514691A)

(43)公表日 令和6年4月2日(2024.4.2)

(51)国際特許分類

F I

G 1 6 B 35/10 (2019.01)

G 1 6 B 35/10

G 1 6 B 40/20 (2019.01)

G 1 6 B 40/20

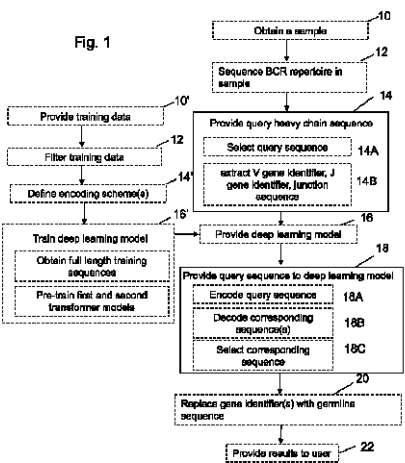
審査請求 未請求 予備審査請求 未請求 (全53頁)			
(21)出願番号	特願2023-564230(P2023-564230)	(71)出願人	523394549
(86)(22)出願日	令和4年4月14日(2022.4.14)		アルケマブ セラピューティクス リミテッド
(85)翻訳文提出日	令和5年12月14日(2023.12.14)		イギリス国, エヌ1シー 4エーエックス
(86)国際出願番号	PCT/EP2022/060073		ロンドン キングス クロス イースト サイド オフィス 1. 0 2
(87)国際公開番号	WO2022/223451	(74)代理人	100079108
(87)国際公開日	令和4年10月27日(2022.10.27)		弁理士 稲葉 良幸
(31)優先権主張番号	2105776.5	(74)代理人	100109346
(32)優先日	令和3年4月22日(2021.4.22)		弁理士 大貫 敏史
(33)優先権主張国・地域又は機関	英国(GB)	(74)代理人	100117189
(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,最終頁に続く	(74)代理人	100134120
			弁理士 江口 昭彦
		(74)代理人	100134120
			弁理士 内藤 和彦
		(72)発明者	リーム, ジヌ
			最終頁に続く

(54)【発明の名称】 抗原結合タンパク質の遺伝子操作

(57)【要約】

【課題】鎖対を含む抗原結合タンパク質を同定する方法が記載される。

【解決手段】方法は、第1鎖配列を含むクエリ配列を提供することと、そのクエリ配列を、入力としてクエリ第1鎖配列をとり、出力として少なくとも1つの対応鎖配列を生成し、それより、クエリ配列の対応鎖配列を同定するように構成された深層学習モデルに提供することにより、対応鎖配列を同定することを含み、深層学習モデルは、既知の鎖対からのトレーニング第1鎖配列及び対応鎖配列を使用してトレーニングされている。第1鎖配列は、抗体若しくはB細胞レセプタの重／軽鎖であってもよく又はT細胞レセプタのβ／α／δ／γ鎖であってもよく、対応鎖は、抗体若しくはB細胞レセプタの軽／重鎖であってもよく又はT細胞レセプタのβ／α／δ／γ鎖であってもよい。方法は、例えば、所望の（例えば、治療又は機能の）性質を有する抗原結合分子を同定する状況等の、抗原結合分子の鎖対を同定することが望ましい任意の状況で利用される。関連する方法、システム、及び製品も記載される。



【特許請求の範囲】

【請求項 1】

鎖対を含む抗原結合タンパク質を同定するコンピュータ実施方法であって、

第 1 鎖配列を含むクエリ配列を提供することと、

入力としてクエリ第 1 鎖配列をとり、出力として少なくとも 1 つの対応鎖配列を生成し、それにより前記クエリ配列の対応鎖配列を同定するように構成された深層学習モデルに前記クエリ配列を提供すること

により、対応鎖配列を同定することを含み、前記深層学習モデルは、既知の鎖対からのトレーニング第 1 鎖配列及び対応鎖配列を使用してトレーニングされている、方法。

【請求項 2】

前記抗原結合タンパク質は、

(i) 重-軽鎖対であって、前記第 1 鎖配列は重鎖配列若しくは軽鎖配列であり、前記対応鎖配列は軽鎖配列若しくは重鎖配列であり、任意選択的に、前記第 1 鎖配列は重鎖配列であり、前記対応配列は軽鎖配列である、重-軽鎖対、又は、

(i i) α β 鎖対であって、前記第 1 鎖配列は β 鎖配列若しくは α 鎖配列であり、前記対応鎖配列は α 鎖配列若しくは β 鎖配列であり、任意選択的に、前記第 1 鎖配列は β 鎖配列であり、前記対応配列は α 鎖配列である、 α β 鎖対、又は、

(i i i) γ δ 鎖対であって、前記第 1 鎖配列は δ 鎖配列若しくは γ 鎖配列であり、前記対応鎖配列は γ 鎖配列若しくは δ 鎖配列であり、任意選択的に、前記第 1 鎖配列は δ 鎖配列であり、前記対応配列は γ 鎖配列である、 γ δ 鎖対

を含む、請求項 1 に記載の方法。

【請求項 3】

前記深層学習モデルは、配列-配列モデルであり、及び／又は前記深層学習モデルは、回帰型ニューラルネットワーク又はトランスフォーマを含む、請求項 1 又は 2 に記載の方法。

【請求項 4】

前記深層学習モデルは、出力として、1 つ又は複数の対応鎖配列を生成するように構成され、任意選択的に、各対応鎖配列には、確率等の信頼メトリックが関連付けられる、請求項 1 ~ 3 のいずれか 1 項に記載の方法。

【請求項 5】

前記深層学習モデルは、配列-配列トランスフォーマベースモデルである、請求項 1 ~ 4 のいずれか 1 項に記載の方法。

【請求項 6】

既知の鎖対からの前記トレーニング第 1 鎖配列及び前記対応鎖配列は、単一 B 細胞シーケンシングデータからの対になったトレーニング重鎖及び軽鎖配列を含み、又は既知の鎖対からの前記トレーニング第 1 鎖配列及び前記対応鎖配列は、単一 T 細胞シーケンシングデータからの対になったトレーニング α 鎖配列及び β 鎖配列を含む、請求項 1 ~ 5 のいずれか 1 項に記載の方法。

【請求項 7】

前記クエリ鎖配列は、V 遺伝子配列又は識別子、J 遺伝子配列又は識別子、及び接合配列、及び任意選択的に D 遺伝子配列又は識別子を含むか又はそれからなり、

前記対応鎖配列は、V 遺伝子配列又は識別子、J 遺伝子配列又は識別子、及び接合配列を含むか又はそれからなる、請求項 1 ~ 6 のいずれか 1 項に記載の方法。

【請求項 8】

前記クエリ鎖配列は、V 遺伝子配列又は識別子、J 遺伝子配列又は識別子、及び接合配列を含むか又はそれからなり、

前記対応鎖配列は、V 遺伝子配列又は識別子、J 遺伝子配列又は識別子、及び接合配列、及び任意選択的に D 遺伝子配列又は識別子を含むか又はそれからなる、請求項 1 ~ 6 のいずれか 1 項に記載の方法。

【請求項 9】

前記クエリ配列は、1つ又は複数の第1鎖CDR配列を含むか又はそれからなり、及び／又は前記対応配列は1つ又は複数の対応鎖CDR配列を含むか又はそれからなり、任意選択的に、前記クエリ／対応配列はCDR3配列を含むか又はそれからなる、請求項1～8のいずれか1項に記載の方法。

【請求項10】

全ての配列はアミノ酸配列である、請求項1～9のいずれか1項に記載の方法。

【請求項11】

前記クエリ配列を前記深層学習モデルに提供することは、エンコード方式を使用して前記クエリ配列をエンコードすることを含み、

各遺伝子配列識別子は個々のトークンに対応し、

各アミノ酸は個々のトークンに対応し、及び／又は

配列は、各々が個々のk-merに対応するトークンを使用して又はバイト対エンコードを使用してエンコードされ、任意選択的に各配列は重複k-merを使用してエンコードされる、請求項1～10のいずれか1項に記載の方法。

【請求項12】

前記対応鎖配列を同定することは、エンコード方式を使用して前記深層学習モデルにより出力された対応配列をデコードすることを含み、

各遺伝子配列識別子は個々のトークンに対応し、

各アミノ酸は個々のトークンに対応し、及び／又は

配列は、各々が個々のk-merに対応するトークンを使用して又はバイト対エンコードを使用してエンコードされ、任意選択的に各配列は重複k-merを使用してエンコードされる、請求項1～11のいずれか1項に記載の方法。

【請求項13】

前記エンコード方式は、前記トレーニング鎖配列の内容に基づいて先に定義されている、請求項11又は12に記載の方法。

【請求項14】

前記クエリ配列及び／又は前記対応配列は、1つ又は複数の遺伝子配列識別子を含み、前記方法は、1つ又は複数の遺伝子配列識別子に対応生殖細胞系配列で置換することを含む、請求項1～13のいずれか1項に記載の方法。

【請求項15】

前記深層学習モデルは、対になっていないトレーニング第1鎖配列及び／又は対応鎖配列を使用して予めトレーニングされたエンコーダと、対になっていないトレーニング対応鎖配列及び／又は第1鎖配列を使用して予めトレーニングされたデコーダとを含む、トランスフォーマベースのモデルである、請求項1～14のいずれか1項に記載の方法。

【請求項16】

前記対になっていないトレーニング鎖配列は、前記対応鎖の可変領域に全長配列を含む、及び／又は前記対になっていないトレーニング鎖配列は、前記第1鎖の前記可変領域に全長配列を含む、請求項15に記載の方法。

【請求項17】

クエリ配列を提供することは、

ユーザインタフェースを通してユーザから、計算デバイスから、配列取得手段又は配列取得手段と関連付けられた計算デバイスから、データベース又は他のコンピュータ可読媒体から、前記クエリ配列を取得すること、

前記クエリ配列を含む抗原結合分子の遺伝物質エンコードを含む試料をシーケンシングすることであって、任意選択的に、前記クエリ配列を取得することは、B細胞を含む試料のB細胞バルクシーケンシング、T細胞を含む試料のT細胞バルクシーケンシング、又は、前記クエリ配列を含む抗原結合分子を発現する任意の他の細胞、若しくは、B細胞レセプタライブラリ若しくはT細胞レセプタライブラリ等のそこから導出された遺伝物質、を含む試料のバルクシーケンシング、を実行することを含む、シーケンシングすること、及び／又は

10

20

30

40

50

B細胞、T細胞、前記クエリ配列を含む抗原結合分子を発現する他の細胞、又はB細胞レセプタライブラリ若しくはT細胞レセプタライブラリ等のそこから導出された遺伝物質、を含む試料を取得すること、を含む、請求項1～16のいずれか1項に記載の方法。

【請求項18】

ユーザインタフェースを通して、前記同定された対応配列、その一部、又はそこから導出された情報をユーザに提供することをさらに含む、請求項1～17のいずれか1項に記載の方法。

【請求項19】

第1鎖配列を含む複数のクエリ配列に抗原結合タンパク質鎖対を提供する方法であって、請求項1～18のいずれか1項に記載の方法を前記クエリ配列の各々に対して実行することを含み、任意選択的に、前記複数のクエリ配列は、バルクB細胞レパートリーシーケンシングにより取得された重又は軽鎖配列である、方法。

10

【請求項20】

所望の性質を有する抗原結合タンパク質を提供する方法であって、

第1鎖配列を含む1つ又は複数のクエリ配列を提供することであって、前記1つ又は複数のクエリ配列の少なくとも1つは前記所望のレパートリーを有する可能性が高い、提供することと、

請求項1～18のいずれか1項に記載の方法を使用して、前記1つ又は複数のクエリ配列の各々の対応鎖配列を同定することと、

20

を含む、方法。

【請求項21】

前記クエリ配列及びそれから導出された1つ又は複数の前記対応配列の一方を各々が含む1つ又は複数の抗原結合タンパク質候補を取得することと、任意選択的に、前記所望の性質について前記1つ又は複数の抗原結合タンパク質候補を検定することとをさらに含む請求項20に記載の方法。

【請求項22】

鎖対を含む抗原結合タンパク質を同定するためのツールを提供する方法であって、

既知の第1鎖及び対応鎖の対からのトレーニング第1配列及び対応配列を含むトレーニングデータを提供することと、

30

前記トレーニングデータを使用して、入力としてクエリ第1鎖配列をとり、少なくとも1つの対応鎖配列を出力として生成するように、深層学習モデルをトレーニングすることと、

を含む、方法。

【請求項23】

前記トレーニング第1鎖配列をエンコードするためのボキャブラリ及び前記トレーニング対応鎖配列をエンコードするためのボキャブラリを取得することをさらに含み、任意選択的に、前記ボキャブラリはエンコード方式を使用して取得され、ここで任意の遺伝子配列識別子は個々のトークンに対応し、及び／又は任意の配列は、少なくとも部分的に、各々が個々のアミノ酸に対応するトークンを使用してエンコードされ、及び／又は任意の配列は、各々が個々のk-merに対応するトークンを少なくとも部分的に使用して、又はバイト対エンコードを使用してエンコードされる、請求項22に記載の方法。

40

【請求項24】

前記トレーニングされた深層学習モデルをユーザに提供することをさらに含む請求項22又は23に記載の方法

【請求項25】

プロセッサと、

命令を含むコンピュータ可読媒体であって、前記命令は、前記プロセッサにより実行されると、前記プロセッサに請求項1～24のいずれか1項に記載の方法のステップを実行させる、コンピュータ可読媒体と

50

を含む、システム。

【請求項 26】

命令を含む 1 つ又は複数のコンピュータ可読媒体であって、前記命令は、1 つ又は複数のプロセッサにより実行されると、前記 1 つ又は複数のプロセッサに請求項 1 ～ 24 のいずれか 1 項に記載の方法のステップを実行させる、1 つ又は複数のコンピュータ可読媒体。

【発明の詳細な説明】

【技術分野】

【0001】

発明の分野

本発明は、入力された重鎖若しくは軽鎖からの重－軽鎖対又は入力された α 鎖若しくは β 鎖からの α － β 鎖対等の単一の入力された可変鎖からの可変鎖対を同定することにより、B 細胞レセプタ、抗体、及び T 細胞レセプタ等の抗原結合タンパク質を遺伝子操作する方法に関する。本発明は、入力された可変鎖、例えば、B 細胞レセプタ／抗体重又は軽鎖から導出される治療用抗体等の抗原結合タンパク質を提供する方法にも関する。

【背景技術】

【0002】

発明の背景

効果的な液性免疫は、B 細胞レセプタ (BCR) を通して異なる抗原に結合可能な多種多様な B 細胞を必要とする。人間の BCR レパートリーの理論上の全体サイズは最高で約 10^{15} 個のバリエーションであると推定され、そのうちの約 10^9 個が随時、1 人の個人内で循環している [Rees, 2020]。BCR は 2 つのタンパク質の 2 つの対で構成される：2 つの重鎖及び 2 つの軽鎖。各 B 細胞は、重鎖及び軽鎖の対（固有である可能性が高い）を発現して、BCR を形成し、BCR は B 細胞の表面上に発現するか、又は抗体として分泌される。6 億を超える異なるヒト重鎖配列及び概ね 7 千万の軽鎖配列が現在、Observed Antibody Space の目録に載っている [Kovaltsuk et al., 2018]。個人の BCR アンサンブル（個人の BCR レパートリーとも呼ばれる）の特徴付けは、種々の疾患の生物学を理解し [Vander Heiden et al., 2017; Bashford-Rogers et al., 2019; Nielsen et al., 2020; Simonich et al., 2019]、新規の治療法抗体薬を発見する [Krawczyk et al., 2019; Galson et al., 2020] ために価値あるツールであることが証明されている。

【0003】

個人の BCR レパートリーを特徴付ける 2 つの主な手法がある：単一 B 細胞シーケンシング及びバルク B 細胞集団のシーケンシングである。単一細胞シーケンシングは、ペアリング情報を重鎖と軽鎖との間で保持するため、抗体発見用途により一般に採用されている。しかしながら、単一細胞シーケンシングはスループットが限られ、異なるプラットフォーム及びプロトコルによる、単一の試料内に存在する BCR レパートリーのカバレッジは様々である。最も高度なマイクロ流体システムであっても、典型的には、1 試料当たり約 10^4 個の B 細胞の配列を回収することしかできない [King et al., 2021; Eccles et al., 2020; Setliff et al., 2019]。ヒトは典型的には、血液 1 ミリリットル当たり約 10^6 個の B 細胞を有し [Mora and Walczak, 2019]、単一細胞手法が、小さな試料であっても完全な B 細胞多様性を特徴付けることができないことを意味する。さらに、単一細胞シーケンシングは非常に特異的な試料要件を有し（例えば、細胞は典型的には、処理されるまで生きたままである必要があり、したがって、新鮮な試料を収集日に処理する必要があり、又は特定のプロトコルに従って凍結された試料を必要とする）、バルクシーケンシングと比べて 1 試料当たりで非常に高いコストを要し（単一細胞シーケンシングは、バルクシーケンシングよりも少なくとも 1 桁高価である）、専用の研究所機器を必要とする。

【0004】

バルク B 細胞集団のシーケンシングは、1 試料当たり約 10^7 個の B 細胞配列をより容

10

20

30

40

50

易に回収することができ [Briney et al., 2019]、これは個人で予期される多様性にかかなり近い。しかしながら、B細胞はライブラリ準備中に溶解するため、重鎖-軽鎖対情報は保存されない。重鎖は抗原結合において支配的な役割を果たし、軽鎖レパートリーよりもはるかに多様である [Kovaltsuk et al., 2018] ため、典型的には、これらのバルク BCRシーケンシング手法は重鎖のみにフォーカスする。しかしながら、抗体発見の場合、抗体の重鎖と軽鎖とを合成させ、機能の特徴付けるように、抗体の重鎖及び軽鎖の両方を有する必要がある。軽鎖対情報におけるギャップは、計算ペアリング方法の開発を促した [Reddy et al., 2020、Zhu et al., 2013、Raybould et al., 2021、Rakocovic et al., 2021]。しかしながら、これらは特定のデータセット及びこれらのデータセット内の少数の特定の配列に限られる。

10

【0005】

同様に、細胞性免疫も、T細胞レセプタ (TCR) を通して異なる抗原に結合可能な多種多様なT細胞を必要とする。人間のTCRレパートリーの全体サイズは最高で約 10^{15} 個の独自の $\alpha\beta$ T細胞レセプタ (TCR) 対で構成されると推定されている [Carter et al., 2019]。対になった $\alpha\beta$ TCRシーケンシングの実験的手法が開発されている (単一細胞手法 [Zheng et al., 2017] 及び多細胞デコンボリューションベースの手法 [Howie et al., 2015] を含む) が、これらは専用のままであり、スループットが限られたままである。したがって、利用可能なTCRレパートリー知識の大部分は、単鎖レパートリー、大部分は β 鎖レパートリーでのバルクシーケンシングに基づく。これは、特に α 及び β TCR鎖が両方ともアロ反応性及び抗原特異性に関わることが示されている [Carter et al., 2019] ため、本質的に制限される。

20

【発明の概要】

【発明が解決しようとする課題】

【0006】

したがって、BCR重鎖-軽鎖対又はTCR $\alpha\beta$ 鎖対等の鎖対を、このペアリング情報を含まないデータから同定する改良された方法がなお必要とされている。

【課題を解決するための手段】

【0007】

発明の概要

BCR重鎖-軽鎖対を同定する問題は、些細とはほど遠い。実際に、BCRレパートリーの多様性により、探索空間は大きくなる。さらに、幾つかの重鎖-軽鎖組合せは安定したBCRをもたらす (ペアリングがランダムであり得ると推測させた観測 [Glanville et al., 2009; Jayaram et al., 2012; DeKosky et al., 2016]) が、限られた数のペアリングしか、標的抗原に結合可能な機能的なBCRを生成しない [Teplyakov et al., 2016; Ling et al., 2018]。これは、機能的なペアリングが非ランダムであり、機能的なペアリングの決定が、非機能的であるにもかかわらず安定し得るペアリングの数により不明瞭になることを示す。実際に、安定した対を予測することができる場合であっても、これらは、実験による検証を要するとともに主に安定性に基づいて選択されたときに検証結果がよくないと予期される、相当数の解を生み出すため、正しい軽鎖の発見が難しいことを意味する。

30

40

【0008】

複数の異なる計算手法が示唆されており、それらの各々は幾つかの大きな欠点がある。第1の手法は、独立してシーケンシングする際、BCRの重鎖の相対頻度と軽鎖の相対頻度をマッチングすることに基づくものであった [Reddy et al., 2010]。この研究では、マウスにまず免疫性が与えられて、強い免疫反応を生み出し、次いで上位4~5個の最も頻度の高い重鎖及び軽鎖をペアリングに選んだ。これらの上位4~5個の配列では、相対頻度に基づくペアリングは可能ではなかった。より近年、Rakocovic et al. [2021]は、試料が少数の高頻度B細胞優勢の場合しかこの手法が機能しないことを示した。Zhu et al. [2013]は、系統発生的ペアリングと呼ばれる方法を提案したが、これは、重鎖及び軽鎖配列データから生成された系統発生ツリーのアーキテクチャを比較することを

50

含む。この方法は、特定のクローン性増殖；この場合、BCRレパートリー全体ではなく既知の抗ウイルス性抗体系統の調査に限定される。Raybould et al. [2021]は、重鎖及び軽鎖の構造モデルを *in silico* でペアリングすることに基づく手法を提案した。手法は本質的に、高品質の構造鑄型の、制限され且つ高度にスキューした可用性により制限され、せいぜい、必ずしも機能性に繋がるわけではない安定性に関連する特徴を同定することができるだけである。さらに、手法は、特定の配列ではなく、同様の配列のファミリーをペアリングすることしかできない（実用的適用性－実験で検証されていない－を制限する）。したがって、本発明者らは、計算による重－軽鎖対の現行の方法が、特定のデータセット及びこれらのデータセット内の配列のみに適用されるという点で制限されることを同定した。実際、存在する検証された手法は、重鎖配列及び軽鎖配列の両方が試料から利用可能であり、データが大きなクローン性増殖優勢であるデータセットにしか適用可能ではなく、これらのデータセット内の限られた数の配列のペアリングしか促進しない。

10

20

30

40

50

【0009】

本発明者らは、抗体発見への一般化された適用では、任意の所与の重鎖に実行可能な軽鎖を生成可能なことが望ましいことをさらに同定した。BCRレパートリーバルクシーケンシングの取り組みが多くの場合、軽鎖よりも重要な機能的役割を果たすと考えられている重鎖のシーケンシングについての限られたリソースにフォーカスするため、重鎖情報のみを使用してこれを生成可能なことがさらに望ましい。これらの問題に取り組むために、本発明者らは、自然言語処理（NLP）における近年の進歩によりインスパイアされた深層学習方法を使用可能であることを前提とした。特に、トランスフォーマ [Vaswani et al., 2017] 及び他の代替の深層学習アーキテクチャを使用して、ニューラル機械翻訳（NMT）問題としてネイティブ軽鎖配列を同定する問題を表現した。トランスフォーマは、広範囲のNLPタスクで現行技術水準の結果を示した [Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2019; Rothe et al., 2020]。したがって、本発明者らは、入力としてBCR重鎖のみを使用して軽鎖を生成するトランスフォーマモデル（「Matchmaker」と呼ばれる）を使用する方法を考案した。既知のペアリングを有する単一細胞データセットの多くのブラインドテストで、それらは、Matchmakerが他の4つの方法よりも正確な軽鎖遺伝子アノテーションを提供することを示した。次いで、既知の標的を有する9つの治療用抗体と一般に罹患した個人に見られるCOVID-19患者からの18個の重鎖とのセットでの予測が検証された。全てのMatchmaker生成抗体は発現に成功し、安定していた；加えて、治療用抗体についての9つの予測のうちの3つ及びCOVID-19抗体についての18の予測のうちの9つが、標的抗原に *in vitro* で結合する兆候を示した。この手法が、タンデムトランスフォーマを使用して、事前トレーニングステップにおいて単鎖配列の「言語を学習」することにより拡張することができることをさらに示した。この手法は、軽鎖対に対する新規の解決策及びバルク重鎖シーケンシングのギャップを埋める道を提供する。本発明者らは、同じ手法を使用して、 $\alpha\beta$ TCR鎖対の問題を解くことができることをさらに同定した。

【0010】

したがって、第1の態様によれば、鎖対を含む抗原結合タンパク質を同定する方法が提供され、本方法は、第1鎖配列を含むクエリ配列を提供することと、入力としてクエリ第1鎖配列をとり、出力として少なくとも1つの対応鎖配列を生成し、それによりクエリ配列の対応鎖配列を同定するように構成された深層学習モデルにクエリ配列を提供することにより、対応鎖配列を同定することを含み、深層学習モデルは、既知の鎖対からのトレーニング第1鎖配列及び対応鎖配列を使用してトレーニングされている。本方法は、以下の特徴の1つ又は複数を有し得る。

【0011】

鎖対は「可変鎖」と呼ぶことができる。「既知の鎖対」という表現は、所望の抗原結合機能を示す抗原結合タンパク質又は少なくとも1つの被験者のB細胞若しくはT細胞レパートリーの一部を形成する抗原結合タンパク質に存在することが既知の可変鎖配列の対を

指す。後者は「ネイティブ」鎖対と呼ぶこともできる。

【0012】

抗原結合タンパク質は、重－軽鎖対を含み得、第1鎖配列は重鎖配列若しくは軽鎖配列であり、対応鎖配列は軽鎖配列若しくは重鎖配列である。第1鎖配列は重鎖配列であってよく、対応配列は軽鎖配列であってよい。抗原結合タンパク質はB細胞レセプタ若しくは抗体又はそれらから導出されたタンパク質であってよい。したがって、抗原結合タンパク質は重－軽鎖対を含み得る。クエリ配列は、重鎖配列又は軽鎖配列を含み得る。対応鎖配列は、軽鎖配列又は重鎖配列であってよい。

【0013】

抗原結合タンパク質は $\alpha\beta$ 鎖対を含み得、第1鎖配列は β 鎖配列若しくは α 鎖配列であり、対応鎖配列は α 鎖配列若しくは β 鎖配列である。第1鎖配列は β 鎖配列であってよく、対応配列は α 鎖配列であってよい。抗原結合タンパク質は $\gamma\delta$ 鎖対を含み得、第1鎖配列は δ 鎖配列若しくは γ 鎖配列であり、対応鎖配列は γ 鎖配列若しくは δ 鎖配列である。第1鎖配列は δ 鎖配列であってよく、対応配列は γ 鎖配列であってよい。抗原結合タンパク質は、T細胞レセプタ又はそこから導出されたタンパク質であってよい。したがって、抗原結合タンパク質は、 $\alpha\beta$ 鎖対又は $\gamma\delta$ 鎖対を含み得る。したがって、クエリ配列は、 β 若しくは δ 鎖配列又は α 若しくは γ 鎖配列を含み得る。対応鎖配列は、 α 若しくは γ 鎖配列又は β 若しくは δ 鎖配列であってよい。

【0014】

深層学習モデルは、配列－配列モデルであり得る。深層学習モデルは、回帰型ニューラルネットワーク又はトランスフォーマであり得る。深層学習モデルは、配列－配列トランスフォーマベースモデルであり得る。回帰型ニューラルネットワークは、ゲート付き回帰型ユニット（GRU）ベースのモデル又は長短期記憶（LSTM）モデルであり得る。例えば、GRUベースのモデルは、GRUベースのエンコーダ及びGRUベースのデコーダを含み得る。エンコーダは、例えば、1024の隠れ次元を有する4層双方向GRUであり得る。デコーダは、例えば、1024の隠れ次元を有する4層順方向のみGRUであり得る。トランスフォーマは、アテンションのメカニズムを使用する深層学習モデルである。トランスフォーマベースのモデルは、自己アテンション及びポイント毎の全結合層エンコーダ及びデコーダの両方に使用するアーキテクチャを有するトランスフォーマモデルであり得る。エンコーダ及び／又はデコーダは、4つの同一層のスタックで構成し得る。エンコーダの各層は2つの下位層を有し得る：つまり、マルチヘッド自己アテンション層及び位置毎の全結合フィードフォワードネットワーク層。デコーダの各層は3つの下位層を有し得る：つまり、自己アテンション下位層、エンコーダスタックの出力に対してマルチヘッドアテンションを実行する層及びフィードフォワードネットワーク層。モデルは1024のフィードフォワード次元を有し得る。

【0015】

深層学習モデルは、出力として、1つ又は複数の対応鎖配列を生成するように構成し得る。各対応鎖配列には、確率等の信頼メトリックを関連付け得る。クエリ配列の対応鎖配列は、1つ又は複数の対応鎖配列の中の最高信頼メトリックが関連付けられた1つ又は複数の対応鎖配列の配列として同定し得る。深層学習モデルは、出力として、単一の対応鎖配列を生成するように構成し得る。深層学習モデルは、各鎖を逐次予測するように構成し得る。換言すれば、深層学習モデルは、欲張り様式で予測を提供するように構成し得る。深層学習モデルは、ビームサーチ手法又はビームスタックサーチ [Zhou & Hansen, 2005] 及び深さ優先ビームサーチ [Furcy & Koenig, 2005] 等の関連する手法を使用して鎖を予測するように構成し得る。深層学習モデルは、出力として、複数の対応する鎖配列を生成するように構成し得る。複数の対応鎖配列の各々には、確率等の信頼メトリックを関連付け得る。最高信頼メトリックが関連付けられた単一の対応鎖配列を報告し得る。したがって、本方法は、1つ又は複数の対応鎖配列の中で最高信頼メトリックが関連付けられた複数の対応鎖配列のうちの1つとして、対応鎖配列を同定することを含み得る。代替的には、予測された全ての対応鎖配列を、有利なことには関連付けられた信頼メ

10

20

30

40

50

リックと一緒に報告し得る。代替的には、予測され、1つ又は複数のさらなる基準を満たす全ての対応鎖配列を報告し得る。例えば、予測され、最高信頼メトリックと関連付けられた対応鎖配列から所定の範囲内の信頼メトリックが関連付けられたあらゆる対応鎖配列を報告し得る。

【0016】

既知の鎖対からのトレーニング第1鎖配列及び対応鎖配列は、単一B細胞シーケンシングデータからの対になったトレーニング重鎖配列及び軽鎖配列を含み得る。トレーニングデータは、被験者から取得された試料の単一B細胞シーケンシング又はそこから導出されたライブラリのシーケンシングにより、各々前に取得された1つ又は複数のデータセットを含み得る。トレーニングデータは、既知の抗体／B細胞レパートリーからの対になったトレーニング重鎖配列及び軽鎖配列をさらに含み得る。例えば、トレーニングデータは、1つ又は複数の抗体／BCRデータベースから、1つ又は複数の既知の治療用抗体／BCRから、及び／又は所望の結合機能を有することが既知の1つ又は複数の抗体／BCRからの対になったトレーニング重鎖配列及び軽鎖配列を含み得る。

10

【0017】

トレーニングデータは、ナイーブB細胞レセプタライブラリからの対になったトレーニング重鎖配列及び軽鎖配列を含み得る。トレーニングデータは、抗原経験B細胞レセプタライブラリからの対になったトレーニング重鎖配列及び軽鎖配列を含み得る。したがって、トレーニングデータは、1つ又は複数の特定の抗原に暴露されたことがある被験者から取得された対になったトレーニング重鎖配列及び軽鎖配列を含み得る。

20

【0018】

既知の鎖対からのトレーニング第1鎖配列及び対応鎖配列は、単一T細胞シーケンシングデータからの対になったトレーニングα鎖配列及びβ鎖配列を含み得る。トレーニングデータは、被験者から取得された試料の単一T細胞シーケンシング又はそこから導出されたライブラリのシーケンシングにより各々前に取得された1つ又は複数のデータセットを含み得る。トレーニングデータは、既知のT細胞レパートリーからの対になったトレーニング第1鎖配列及び対応鎖配列をさらに含み得る。例えば、トレーニングデータは、1つ又は複数のT細胞レセプタデータベースから、1つ又は複数の既知の治療用TCRから、及び／又は所望の結合機能を有することが既知の1つ又は複数のTCRからの対になったトレーニングα鎖配列及びβ鎖配列を含み得る。トレーニングデータは、ナイーブT細胞レセプタライブラリからの対になったトレーニングα鎖配列及びβ鎖配列（又はδ及びγ）を含み得る。トレーニングデータは、抗原経験T細胞レセプタライブラリからの対になったトレーニングα鎖配列及びβ鎖配列（又はδ鎖配列及びγ鎖配列）を含み得る。したがって、トレーニングデータは、1つ又は複数の特定の抗原に暴露されたことがある被験者から取得された対になったトレーニングα鎖配列及びβ鎖配列（又はδ鎖配列及びγ鎖配列）を含み得る。

30

【0019】

既知の鎖対からのトレーニング第1鎖配列及び対応鎖配列は、対になったトレーニング鎖配列を含み得、各対は、V遺伝子配列又は識別子、J遺伝子配列又は識別子、及び接合配列、及び任意選択的にD遺伝子配列又は識別子を含むか又はそれからなる鎖配列を含む。既知の鎖対からのトレーニング第1鎖配列及び対応鎖配列は、対になったトレーニング鎖配列を含み得、各対は、V遺伝子配列又は識別子、J遺伝子配列又は識別子、及び接合配列を含むか又はそれからなる鎖配列を含むか又はそれからなる鎖配列を含む。トレーニングデータは、少なくとも80,000、少なくとも100,000、少なくとも120,000、又は少なくとも150,000対のトレーニング配列、例えば、トレーニング重鎖配列及び計算配列を含み得る。有利なことには、トレーニングデータは、少なくとも150,000対のトレーニング重鎖配列及び軽鎖配列を含み得る。トレーニングデータは、哺乳類、例えばヒト等の鎖配列対を含み得る。トレーニングデータは、哺乳類の重鎖配列及び／又は軽鎖配列を含み得る。トレーニングデータは、ヒトの重鎖配列及び／又は軽鎖配列を含み得る。トレーニングデータは、クエリ配列と同じ種からのトレーニング配

40

50

列対を含み得る。トレーニングデータは、クエリ配列と同じ種からの少なくとも50%、少なくとも70%、少なくとも80%、少なくとも90%、又は少なくとも95%の配列を含み得る。クエリ配列は、トレーニングデータに存在しない配列であり得る。クエリ配列は、所望の表現型等の所望の特性を有する被験者からの試料から取得された配列であり得る。例えば、被験者は特定の臨床特性を有し得る。

【0020】

トレーニングデータは、対になっていないトレーニング第1配列及び／又は対応配列をさらに含み得る。これについてさらに以下に説明する。対になっていないトレーニング第1鎖配列及び／又は対応鎖配列は、対になった配列に関連して説明した配列の任意の特徴を有し得る。特に、対になっていない鎖配列は、対になった配列と同じタイプの配列であってもよく（例えば、対になったトレーニング配列が重鎖及び軽鎖対である場合、対になっていないトレーニング第1鎖配列／対応鎖配列は対になっていない重鎖及び／又は軽鎖を含み得る）、同じ有機体からの配列を含んでもよく（例えば、哺乳類の配列及び／又はヒトの配列を含んでもよく、1つ又は複数の有機体からの配列を含んでもよく、ナイーブライブラリ及び／又は抗原暴露ライブラリ等からの配列を含んでもよい）、同じ情報（例えば、遺伝子分節識別子、配列、及びそれらの組合せ等）を含んでもよい。対になっていないトレーニング配列は、対になったトレーニング配列に存在する第1配列及び／又は対応配列の幾つか又は全てを含み得る。有利なことには、対になっていないトレーニング配列は、対になったトレーニング鎖配列よりも多くの第1鎖配列及び／又はより多くの対応鎖配列を含み得る。追加の対になっていないトレーニング配列の使用は、深層学習モデルが、対になっていないトレーニング第1配列及び／又は軽対応配列を使用して予めトレーニングされたエンコーダと、対になっていないトレーニング対応鎖配列及び／又は第1鎖配列を使用して予めトレーニングされたデコーダとを含むトランスフォーマベースのモデルである実施形態において特に有利であり得る。

【0021】

クエリ鎖配列は、V遺伝子配列又は識別子、J遺伝子配列又は識別子、及び接合配列、及び任意選択的にD遺伝子配列又は識別子を含み得るか又はそれからなり得る。対応鎖配列は、V遺伝子配列又は識別子、J遺伝子配列又は識別子、及び接合配列を含み得るか又はそれからなり得る。クエリ鎖配列は、V遺伝子配列又は識別子、J遺伝子配列又は識別子、及び接合配列を含み得るか又はそれからなり得る。対応鎖配列は、V遺伝子配列又は識別子、J遺伝子配列又は識別子、及び接合配列、及び任意選択的にD遺伝子配列又は識別子を含むか又はそれからなる。クエリ鎖配列及び対応鎖配列の形式は、トレーニング鎖配列の形式に関連する。したがって、V遺伝子配列又は識別子、J遺伝子配列又は識別子、及び接合配列、及び任意選択的にD遺伝子配列又は識別子を含むか又はそれからなるトレーニング鎖配列を使用してトレーニングされた深層学習モデルは、これらの成分を含むか又はこれらの成分からなる鎖配列を入力として受け入れ得又は出力として生成し得る。同様に、V遺伝子配列又は識別子、J遺伝子配列又は識別子、及び接合配列を含むか又はそれからなるトレーニング鎖配列を使用してトレーニングされた深層学習モデルは、これらの成分を含むか又はこれらの成分からなる鎖配列を入力として受け入れ得又は出力として生成し得る。クエリ配列は、1つ若しくは複数の第1鎖CDR配列を含み得又は1つ若しくは複数の第1鎖CDR配列からなり得る。対応配列は、1つ若しくは複数の対応鎖CDR配列を含み得又は1つ若しくは複数の対応鎖CDR配列からなり得、任意選択的に、クエリ／対応配列はCDR3配列を含むか又はCDR3配列からなる。

【0022】

全ての配列はアミノ酸配列であり得る。クエリ配列を深層学習モデルに提供することは、エンコード方式を使用してクエリ配列をエンコードすることを含み得、各遺伝子配列識別子は個々のトークンに対応する。クエリ配列を深層学習モデルに提供することは、エンコード方式を使用してクエリ配列をエンコードすることを含み得、各アミノ酸は個々のトークンに対応する。クエリ配列を深層学習モデルに提供することは、エンコード方式を使用してクエリ配列をエンコードすることを含み得、配列（即ち、遺伝子識別子ではなく完

全な配列として利用可能な配列)は、各々が個々の k -mer に対応するトークンを使用して又はバイト対エンコードを使用してエンコードされる。各配列は、重複 k -mer (overlapping k -mers) を使用してエンコードし得る。 k -mer は長さ 1~5 であり得る。 k -mer は固定長であり得る。例えば、1、2、3、4、又は 5 の固定 k -mer 長が使用され得る。長さ 1 の k -mer は、各キャラクタ (例えば、各アミノ酸) を個々にエンコードすることと等しい。長さ $k > 2$ (例えば 3 等) の k -mer は、重複又は非重複 k -mer を使用するエンコード方式の一環として使用され得る。重複 k -mer は、異なる程度で重複し得る。例えば、長さ 3 の k -mer は、1 又は 2 キャラクタ重複し得る。 $k = 3$ を使用する方式では、各トークンは 3 キャラクタの独自のセット (例えば、3 つのアミノ酸のモチーフ) に対応する。対応する鎖配列の同定は、各遺伝子配列識別子が個々のトークンに対応するエンコード方式を使用して、深層学習モデルにより出力された対応配列をデコードすることを含み得る。対応する鎖配列の同定は、各アミノ酸が個々のトークンに対応するエンコード方式を使用して、深層学習モデルにより出力された対応配列をデコードすることを含み得る。対応する鎖配列の同定は、配列が、個々の k -mer に各々対応するトークンを使用して又はバイト対エンコードを使用してエンコードされるエンコード方式を使用して、深層学習モデルにより出力された対応配列をデコードすることを含み得る。各配列は、重複 k -mer を使用してエンコードし得る。エンコード方式は、トレーニング鎖配列の内容に基づいて先に定義されていることがある。エンコード方式は、トレーニング鎖配列の内容に基づいて先に定義されていることがあり、トークンは、トレーニングデータにおいて所定の閾値 (例えば、2) を下回る回数、使用された場合 (対になったトレーニングデータにおける第 1 鎖配列及び対応鎖配列で別個に又はまとめて)、トレーニング鎖配列の内容に基づいて構築されるボキャブラリから除外される。エンコード方式は、トレーニングデータにおけるトレーニング第 1 鎖及びトレーニング対応鎖のボキャブラリを別個に構築することにより、トレーニング鎖配列の内容に基づいて先に定義されていることがある。

10

20

30

40

【0023】

トレーニングデータは、長さの所定の範囲外の接合配列 (第 1 鎖及び／又は対応鎖内に) を含む任意の対を除外するようにフィルタリングされていてよい。換言すれば、トレーニングデータは、長さの所定の範囲外の第 1 (例えば、重) 鎖接合及び／又は長さの所定の範囲外の対応 (例えば、軽) 鎖接合を含むいかなる対も含まなくてよい。例えば、アミノ酸 3 個、4 個、5 個、6 個、7 個、8 個、9 個、又は 10 個等の所定の長さを下回る重鎖接合配列を含む対は除外し得る。別の例として、例えば、アミノ鎖 25 個、26 個、27 個、28 個、29 個、30 個、31 個、32 個、33 個、34 個、又は 35 個等の所定の長さを超える重鎖接合配列を含む対は除外し得る。別の例として、例えば、アミノ酸 3 個、4 個、5 個、6 個、7 個、8 個、9 個、又は 10 個等の所定の長さを下回る軽鎖接合配列を含む対は除外し得る。別の例として、例えば、アミノ鎖 15 個、16 個、17 個、18 個、19 個、20 個、21 個、22 個、23 個、24 個、25 個、26 個、27 個、28 個、29 個、又は 30 個等の所定の長さを超える軽鎖接合配列を含む対は除外し得る。所定の長さは、対の対応 (例えば、軽) 鎖及び第 1 (例えば、重) 鎖における接合配列で同じであってもよく又は異なってもよい。具体例では、アミノ酸 7 個よりも少ない重鎖接合配列を含む対は除外し得及び／又はアミノ酸 30 個を超える重鎖接合配列を含む対は除外し得る。これに代えて又は加えて、アミノ酸 7 個よりも少ない軽鎖接合配列を含む対は除外し得及び／又はアミノ酸 20 個を超える軽鎖接合配列を含む対は除外し得る。

【0024】

クエリ配列及び／又は対応配列は、1 つ又は複数の遺伝子配列識別子を含み得、本方法は、1 つ又は複数の遺伝子配列識別子に対応生殖細胞系配列で置換することをさらに含み得る。

【0025】

深層学習モデルは、対になっていないトレーニング第 1 鎖配列及び／又は対応鎖配列を

50

使用して予めトレーニングされたエンコーダと、対になっていないトレーニング対応鎖配列及び／又は第1鎖配列を使用して予めトレーニングされたデコーダとを含むトランスフォーマベースのモデルであり得る。エンコーダ及び／又はデコーダは、BERTモデル、若しくは例えば、BERT、RoBERTa、又はDistilBERT等のその変形、及び／又はGPT-2等の自己回帰型トランスフォーマモデルを含み得る。エンコーダ及び／又はデコーダは、RoBERTaモデル、BERTモデル、及び／又はGPT-2モデルを含み得る。エンコーダ及びデコーダは同じモデルを含み得る。エンコーダ及びデコーダは両方とも、対になっていないトレーニング対応鎖配列及び第1鎖配列を使用してトレーニングされたモデルを含み得る。例えば、エンコーダ及びデコーダは両方とも、各々第1鎖配列及び対応鎖配列を含むランダム対を使用してトレーニングされたモデルを含み得る。代替的には、エンコーダはトレーニングされた第1（例えば、重又は軽）鎖配列を使用してトレーニングされたモデルを含み得、デコーダは対応（例えば、軽又は重）鎖配列を使用してトレーニングされたモデルを含み得る。エンコーダ及びデコーダが両方とも、対になっていないトレーニング第1鎖配列及び対応鎖配列を使用してトレーニングされたモデルを含む場合、エンコーダ及びデコーダは両方とも同じ予めトレーニングされたモデルを含み得る。したがって、エンコーダ及びデコーダは、同じパラメータと共に同じアーキテクチャを有する、予めトレーニングされたモデルを使用して初期化し得る。対になっていないトレーニング鎖配列は、対応鎖の可変領域に全長配列を含み得る。対になっていないトレーニング対応鎖配列は、第1鎖の可変領域に全長配列を含み得る。トランスフォーマベースのモデルは、既知の鎖対からの対になった第1鎖配列及び対応（例えば、重及び軽）鎖配列を使用してトレーニングされた可能性があり、上記配列は、対応鎖及び／又は第1鎖の可変領域に完全な長さの配列を含まない。そのような実施形態では、トランスフォーマベースのモデルは、欠損配列情報を推定することで対応鎖及び／又は第1鎖の可変領域に完全な長さの配列を含む対になったトレーニング配列を取得することによりトレーニングされていてよい。

10

20

【0026】

欠損配列情報をインプットすることは、対応生殖細胞系配列で遺伝子識別子を置換することを含み得る。欠損配列情報をインプットすることは、予めトレーニングされたエンコーダ及び／又は予めトレーニングされたデコーダを使用して（例えば、欠損配列情報が推定される鎖に応じて）、対になったトレーニング第1（例えば、重）鎖配列及び／又は対応（例えば、軽）鎖配列の各々から全長配列を予測することを含み得る。代替的には、対になっていないトレーニング対応（例えば、軽）鎖配列及び／又は対になっていないトレーニング第1（例えば、重）鎖配列は、エンコーダ及び／又はデコーダをトレーニングする前、対になった各トレーニング配列の形式に一致する形式に変換されていることがある。

30

【0027】

クエリ配列を提供することは、ユーザインタフェースを通してユーザから、計算デバイスから、配列取得手段又は配列取得手段と関連付けられた計算デバイスから、データベース又は他のコンピュータ可読媒体からクエリ配列を取得することを含み得る。クエリ配列を提供することは、クエリ配列を含む抗原結合分子の遺伝物質エンコードを含む試料をシーケンシングすることを含み得る。クエリ配列を取得することは、B細胞を含む試料のB細胞バルクシーケンシング、T細胞を含む試料のT細胞バルクシーケンシング、又はクエリ配列を含む抗原結合分子を発現する任意の他の細胞若しくはB細胞レセプタライブラリ若しくはT細胞レセプタライブラリ等のそこから導出された遺伝物質を含む試料のバルクシーケンシングを実行することを含み得る。クエリ配列を提供することは、B細胞、T細胞、クエリ配列を含む抗原結合分子を発現する他の細胞、又はB細胞レセプタライブラリ若しくはT細胞レセプタライブラリ等のそこから導出された遺伝物質を含む試料を取得することを含み得る。クエリ配列を提供することは、例えば、B細胞（又はクエリ配列を含む抗原結合分子を発現している任意の他の細胞又はB細胞レセプタライブラリ等のそこから導出される遺伝物質）を含む試料のB細胞バルクシーケンシングを実行することにより

40

50

、クエリ配列を含む抗原結合分子の遺伝物質エンコードを含む試料をシーケンシングすることを含み得る。クエリ配列を提供することは、B細胞、又はクエリ配列を含む抗原結合分子を発現している他の細胞、又はB細胞レセプタライブラリ等のそこから導出された遺伝物質を含む試料を取得することを含み得る。

【0028】

本方法は、ユーザインタフェースを通して、同定された対応配列、その一部、又はそこから導出された情報をユーザに提供することをさらに含み得る。

【0029】

第2の態様によれば、第1鎖配列を含む複数のクエリ配列に抗原結合タンパク質鎖対を提供する方法が提供され、本方法は、各クエリ配列に対して第1の態様の任意の実施形態の方法を実行することを含む。複数のクエリ配列は、バルクB細胞レパートリーシーケンシングにより取得された重又は軽鎖配列であり得る。複数のクエリ配列は、少なくとも100、少なくとも1000、少なくとも10,000、又は少なくとも100,000の配列を含み得る。複数のクエリ配列は、被験者からの試料等の試料中の重又は軽鎖レパートリーのバルクB細胞シーケンシングにより取得されていてよい。複数の配列は、試料中の重又は軽鎖レパートリーのバルクB細胞シーケンシングにより取得された配列のセットのうちのサブセットであり得る。本発明の態様による本方法は、第1の態様に関連して説明した任意の特徴を有し得る。

【0030】

第3の態様によれば、所望の性質を有する抗原結合タンパク質を提供する方法が提供され、本方法は、第1鎖配列を含む1つ又は複数のクエリ配列を提供することであって、1つ又は複数のクエリ配列の少なくとも1つは所望の性質を有する可能性が高い、提供することと、第1の態様の任意の実施形態の方法を使用して1つ又は複数のクエリ配列の各々の対応鎖配列を同定することを含む。本方法は、以下の特徴の1つ又は複数を含み得る。

【0031】

本方法は、クエリ配列及びそれから導出された1つ又は複数の対応配列の一方を各々が含む1つ又は複数の抗原結合タンパク質候補を取得することをさらに含み得る。本方法は、所望の性質について1つ又は複数の抗原結合タンパク質候補を検定することをさらに含み得る。本態様の方法は、第1又は第2の態様に関連して説明した任意の特徴を有し得る。1つ又は複数の抗原結合タンパク質候補は、抗体又はその断片であり得る。同定された鎖対から導出された配列は、同じCDRを含むが、フレームワーク領域が異なる配列、同定された鎖対と比べて1つ又は複数の変異を含む配列、及び同定された鎖対の1つ又は複数の断片を含む配列を含み得る。抗原結合タンパク質候補を取得することは、抗原結合タンパク質候補のコード配列を同定することと、適した発現系（例えば、適した宿主細胞等）で配列を発現させることとを含み得る。

【0032】

所望の性質は、所望の結合性質（例えば、1つ若しくは複数の標的に結合する能力、1つ若しくは複数の各閾値を超える親和性をもって1つ若しくは複数の標的に結合する能力等）、所望の発現性質（例えば、1つ若しくは複数の発現系における標準と比べて高い発現レベル、1つ若しくは複数の発現系における所定のレベルを超える発現レベル、1つ若しくは複数の発現系における所定のレベルを超える収率等）、所望の安定性質（例えば、1つ若しくは複数の条件で特定の閾値を超える安定性等）、又はそれらの組合せであり得る。所望の性質は、所定の標的に結合する能力を含み得る。所望の性質について1つ又は複数の抗原結合タンパク質候補を検定することは、例えば、1つ又は複数の抗原候補への結合について検定することにより、1つ又は複数の抗原結合タンパク質候補が結合する1つ又は複数の抗原を同定することを含み得る。所望の性質について1つ又は複数の抗原結合タンパク質候補を検定することは、例えば、既知の標的を有する1つ又は複数の抗体との比較により、1つ又は複数の抗原結合タンパク質候補が結合する1つ又は複数の抗原を同定することを含み得る。抗原結合タンパク質は治療用抗体であり得、所望の性質は治療

10

20

30

40

50

標的の結合を含み得る。抗原結合タンパク質は、本明細書では「免疫タンパク質」と呼ばれることもある。

【0033】

所望の性質について1つ又は複数の抗原結合タンパク質候補を検定することは、有機体（例えば、動物モデル等）における所望の表現型の有無又は1つ若しくは複数の抗原結合タンパク質候補を発現する細胞の有無を同定することを含み得る。所望の表現型の存在を同定することは、1つ又は複数のモデル細胞（例えば、1つ若しくは複数の細胞株）又は有機体（例えば、1つ若しくは複数の動物モデル等）内で1つ又は複数の抗原結合タンパク質候補を発現させることを含み得る。

【0034】

本方法は、1つ又は複数の抗原結合タンパク質候補の少なくとも1つの配列を最適化することをさらに含み得る。抗原結合タンパク質候補の配列を最適化することは、例えば、当技術分野において既知の任意の抗体最適化技法を使用して実行し得る。抗原結合タンパク質候補の配列を最適化することは、例えば、鎖対が同定された入力配列と同様の配列を解析することにより、鎖対が同定された配列データからの情報を使用して実行し得る。抗原結合タンパク質を最適化する方法は、当技術分野で既知であり、とりわけ、Mason et al. [2021]、Seeliger et al., [2015]、Warszawski et al. [2019]、Hsiao et al. [2019]、及びRichardson et al. [2021]に記載の方法を含む。これらの方法のいずれも本発明の状況内で使用することが可能である。

【0035】

クエリ配列は、既知の抗体の重鎖配列（又は重鎖配列の一部）を含み得る。したがって、第1鎖は、既知の抗体の重鎖配列又は重鎖配列の一部であり得る。

【0036】

クエリ配列は、1つ又は複数の試料中の重鎖レパートリーのバルクBCRシーケンシングにより取得されていてよい。本方法は、1つ又は複数の試料中の重鎖レパートリーのバルクBCRシーケンシングによりクエリ配列を取得するステップを含み得る。1つ又は複数の試料は、1人又は複数人の被験者からの試料であり得る。1人又は複数人の被験者は、所望の特性、例えば、特定の臨床表現型又はバイオマーカープロファイル等の臨床的に関連する特性等を有するものとして同定されていてよい。例えば、1人又は複数人の被験者は、特定の疾患又は症状に対する回復力を有し得る。疾患又は症状は、がん（例えば、乳がん等）、神経変性疾患（例えば、筋萎縮性側索硬化症等）、及び感染疾患（例えば、COVID-19等）から選択し得る。

【0037】

本方法は、1つ又は複数の試料において同定された第1（例えば、重）鎖配列から選択される複数のクエリ鎖配列（例えば、重鎖配列）の鎖対（例えば、重-軽対）を同定し、それにより、鎖対（例えば、重-軽鎖対）のセットを取得することを含み得る。

【0038】

本方法は、複数のペプチド候補と突き合わせて、1つ又は複数の試料と同じソースからの抗体をスクリーニングすることにより、1つ又は複数の標的を同定することをさらに含み得る。複数のペプチド候補は、1つ又は複数の試料の由来元である種に基づいて選択し得る。例えば、1つ又は複数の試料のソースは、1人又は複数人のヒト被験者であり得、1つ又は複数の試料と同じソースからの抗体レパートリーは、ヒトペプチドームを表すペプチド候補のセットと突き合わせてスクリーニングされて、複数のペプチド候補を選択し得る。

【0039】

1つ又は複数の抗原結合タンパク質候補が結合する抗原を同定することは、複数のペプチド候補と突き合わせて、1つ又は複数の試料と同じソースからの抗体をスクリーニングすることにより同定された1つ又は複数の標的を使用することを含み得る。本方法は、1つ又は複数の基準に基づいて、同定された鎖対のセットをフィルタリングすることをさらに含み得る。1つ又は複数の基準は、抗原結合タンパク質候補が結合するか、又は結合す

10

20

30

40

50

ると予測される抗原又は抗原のセットの同定に適用し得る。

【0040】

1つ又は複数のクエリ配列を提供することは、第1のクエリ（例えば、重）鎖配列及び第2のクエリ（例えば、重）鎖配列を提供することを含み得、1つ又は複数のクエリ配列の各々の対応（例えば、軽）鎖配列を同定することは、1つ又は複数の第1の対応（例えば、軽）鎖配列及び1つ又は複数の第2の対応（例えば、軽）鎖配列を同定することを含み得る。本方法は、第1の対応鎖配列と第2の対応鎖配列とを比較して、第1（例えば、重）鎖の両方を含む二重特異性抗体の共通対応（例えば、軽）鎖としての使用に適し得る1つ又は複数の軽鎖を同定することをさらに含み得る。

【0041】

第4の態様によれば、鎖対を含む抗原結合タンパク質を同定するためのツールを提供する方法が提供され、本方法は、既知の第1鎖及び対応鎖の対からのトレーニング第1配列及び対応配列を含むトレーニングデータを提供することと、トレーニングデータを使用して、入力としてクエリ第1鎖配列をとり、出力として少なくとも1つの対応鎖配列を生成するように深層学習モデルをトレーニングすることを含む。本態様の方法は、第1の態様に関連して説明した任意の特徴を有し得る。本方法は、以下の特徴の任意の1つ又は複数を含み得る。本方法は、トレーニング第1鎖配列をエンコードするためのボキャブラリ及びトレーニング対応鎖配列をエンコードするためのボキャブラリを取得することをさらに含み得る。ボキャブラリは、エンコード方式を使用して取得し得、ここで、任意の遺伝子配列識別子は個々のトークンに対応する。ボキャブラリは、任意の配列が少なくとも部分的に、各々が個々のアミノ酸に対応するトークンを使用してエンコードされるエンコード方式を使用して取得し得る。ボキャブラリは、各々が個々のk-merに対応するトークンを少なくとも部分的に使用して任意の配列がエンコードされ、又はバイト対エンコードを使用してエンコードされるエンコード方式を使用して取得し得る。

【0042】

トレーニングデータを提供することは、対になっていない第1鎖配列及び対応鎖配列を提供することを含み得る。対になっていないトレーニング第1鎖配列及び対応鎖配列はトレーニング前データと呼ぶことができる。本方法は、対になっていないトレーニング第1及び／又は対応（例えば、重及び／又は軽）鎖配列を使用して第1のトランスフォーマベースのモデルをトレーニングし、対になっていないトレーニング対応及び／又は第1（例えば、軽及び／又は重）鎖配列を使用して第2のトランスフォーマベースのモデルをトレーニングすることと、予めトレーニングされた第1及び第2のトランスフォーマモデルを使用して、深層学習モデルのエンコーダ及びデコーダをそれぞれ初期化することとをさらに含み得る。

【0043】

第1及び第2のトランスフォーマベースのモデルは各々、BERTモデル又は例えば、BERT、RoBERTa、若しくはDistilBERT等のその変形又はGPT-2等の自己回帰型トランスフォーマモデルを含み得る。第1及び第2のトランスフォーマベースのモデルは各々、RoBERTaモデル、BERTモデル、又はGPT-2モデルを含み得る。

【0044】

本方法は、トレーニングされた深層学習モデルをユーザに提供することをさらに含み得る。

【0045】

本明細書に記載の方法は、文脈により別段のことが示される場合、例えば、試料が取得、処理、解析され、又は分子若しくは組成物が産生、検定、若しくは任意の他の目的で使用される場合等を除き、コンピュータで実施される。

【0046】

第5の態様によれば、システムが提供され、本システムは、プロセッサと、命令を含むコンピュータ可読媒体であって、命令は、プロセッサにより実行されると、プロセッサに

10

20

30

40

50

任意の先の態様の任意の実施形態の方法のステップを実行させる、コンピュータ可読媒体とを備える。命令は、プロセッサに、第1及び／又は第4の態様の任意の実施形態の方法のステップを実行させ得る。

【0047】

第6の態様によれば、命令を含む1つ又は複数のコンピュータ可読媒体であって、命令は、1つ又は複数のプロセッサにより実行されると、1つ又は複数のプロセッサに、任意の先の態様の任意の実施形態の方法のステップを実行させる、1つ又は複数のコンピュータ可読媒体が提供される。命令は、プロセッサに、第1及び／又は第4の態様の任意の実施形態の方法のステップを実行させ得る。

【図面の簡単な説明】

【0048】

図面の簡単な説明

【図1】本開示による鎖対を同定する方法を概略的に示すフローチャートである。

【図2】本開示による鎖対を同定するシステムの一実施形態を示す。

【図3A】入力として重鎖のみを使用する複数の軽鎖予測方法を概略的に示す。A：重鎖トークン化及び配列への軽鎖トークン変換を用いるトランスフォーマーアーキテクチャ。

【図3B】入力として重鎖のみを使用する複数の軽鎖予測方法を概略的に示す。B：重鎖入力及び軽鎖出力を有するGRUモデル。

【図3C】入力として重鎖のみを使用する複数の軽鎖予測方法を概略的に示す。C：データベース探索方法ワークフロー。

【図3D】入力として重鎖のみを使用する複数の軽鎖予測方法を概略的に示す。D：頻度探索が重及び軽鎖リードカウントのランク付き分布に一致し、次いで、同様のランクを有する鎖をペアリングする。ランダム探索は、データベース探索法の変形であり、示されていない。

【図4】重鎖及び軽鎖のアーキテクチャと、このアーキテクチャに関連して本明細書で使用するデータの構成とを概略的に示す。A．Ig重鎖のアーキテクチャ。V、D、及びJ遺伝子の大きな境界が、接合の境界と共に記される。N末端に向かうV遺伝子の分節は点線境界内にあり、その理由は、多くのNGS法からのリード長が、この領域をカバーするには短すぎ及び／又はNGSに使用される幾つかのプライマーがV領域内にわずかに差し込まれるためである。しかしながら、実線境界内の配列を使用してV遺伝子を推測することがなお可能である。B．Aと同じであるが、軽鎖の場合。C．DeKosky et al. (2015)からの対になったリードアーキテクチャの設計の再作成。

【図5】重鎖入力及び本明細書に記載のトランスフォーマーモデルからの軽鎖予測のアテンションヒートマップを示す。各列は入力重鎖トークンに対応し、各行は出力軽鎖トークンを表す。8つのアテンションヘッドのうちの4つが示され、各ヘッドが重鎖の異なるトークンにいかに関心を持つかを示す。

【図6A】図3に示された方法でのヘルドアウトテストセット及び単一細胞ブラインドテストでの予測性能を示す。A：正しい軽鎖V遺伝子を有する予測の割合。

【図6B】図3に示された方法でのヘルドアウトテストセット及び単一細胞ブラインドテストでの予測性能を示す。B：King et al. [2021]データセットで予測された軽鎖接合配列のレーベンシュタイン距離分布。レーベンシュタイン距離は、予測された軽鎖接合アミノ酸配列がオリジナルと異なるかを定量化するメトリックであり、距離が長いほど不良な予想を意味する。

【図7】ネイティブペルツズマブ軽鎖（黄色）と重鎖（白色）との接触の概略表現を示す；PDB：1s78。画像の中心に、アラニンが球棒模型で示されており、より大きな側鎖を有するアスパラギンは、Matchmakerが予測したように、クラッシュを生じさせ、発現なしに繋がった可能性がある。

【図8】アテゾリズマブ結合のELISAトレースを示す。塗りつぶされた点は標的に対する異なる抗体の結合曲線を表し、一方、塗りつぶされていない点は無関係の抗原に対する結合を表す。

10

20

30

40

50

【図 9】 2つのリンクされた A n t i B E R T a モデルを使用して、本明細書に記載のタンデムトランスフォーマモデルのトレーニング手順を示す。A．マスキングされた言語モデルタスクのセットのトレーニング、検証、及び検定の作成。B．トレーニング前手順のセットアップ及び続くステップに向けてモデルフィードをいかに「ウォームアップ」するか。C．ウォームアップされたモデルを N M T の S e q 2 S e q モデルの一環としていかに使用することができるかの概説。

【発明を実施するための形態】

【0049】

発明の詳細な説明

本発明を説明するに当たり、以下の用語が採用され、以下示されるように定義されることが意図される。

【0050】

B細胞レセプタは、B細胞の表面に発現する膜貫通タンパク質である。B細胞レセプタは、同種抗原を認識する膜結合免疫グロブリン分子（抗体とも呼ばれる）を含む結合部分（「抗原結合サブユニット」又は「膜免疫グロブリン」、「m l g」とも呼ばれる）と、シグナル伝達部分とを含む。膜結合免疫グロブリン分子は、2つの免疫グロブリン軽鎖及び2つの免疫グロブリン重鎖を含み、膜内在性ドメインを除き、対応する分泌される抗体と同一である。シグナル伝達部分は、一緒に結合されるとともに、ジスルフィド架橋により免疫グロブリンに結合された $I g - \alpha / I g - \beta$ (C D 7 9) と呼ばれるヘテロ二量体である。抗体 (A b) 又は免疫グロブリン (I g) は、抗原結合部位と、アイソタイプの限られたセット (I g A、I g D、I g E、I g G、又はI g M) の1つに属し、免疫系の他の成分との相互作用を仲介する定常領域とを含む免疫タンパク質である。ヒト及び大半の哺乳類では、抗体は4つのポリペプチド鎖を含む：ジスルフィド結合により接続された2つの同一の重鎖及び2つの同一の軽鎖。軽鎖は典型的には1つの可変ドメイン V_L 及び1つの定常ドメイン C_L からなり、一方、重鎖は典型的には1つの可変ドメイン V_H 及び3～4つの定常ドメイン $C_H 1$ 、 $C_H 2$ 、・・・からなる。可変ドメインは、抗原結合領域を形成し、 F_v 領域と呼ぶこともできる。各可変ドメインは、一緒になって抗原結合部位を形成する、相補性決定領域 (C D R) と呼ばれる3つの高頻度可変性領域を含む。各免疫グロブリン重鎖又は軽鎖の可変領域は、幾つかの個片—遺伝子分節（小遺伝子）として知られる—にエンコードされる：I g 重鎖は可変 (V) 分節、多様性 (D) 分節、及び接合 (J) 分節を含み、I g 軽鎖はV分節及びJ分節を含む。V、D、及びJ遺伝子分節の複数のコピーがゲノムに存在し、B細胞前駆細胞は、V (D) J 遺伝子再構成と呼ばれるプロセスで1つのV、1つのD、及び1つのJ遺伝子分節（又は軽鎖において1つのV及び1つのV分節）を（略）ランダムに選択し組み合わせることにより、I g 可変領域を組み立てる。このプロセスは、必要とされる分節間に、ヘアピンループを形成する2本ストランド切断を形成することを含み、ヘアピンループは次いで一緒に接合される。接合プロセスは不正確であり、V分節とJ分節との間（軽鎖）又はV分節とD J分節とD分節とJ分節との間（重鎖）にヌクレオチドの可変追加又は可変削減が生じ、分節間の接合の配列（「接合配列」と呼ばれる）に大きな多様性を生み出す。V (D) J 遺伝子再構成プロセスは、I g の抗原結合領域に新規のアミノ鎖配列を産生し、抗原認識能力の膨大な多様性を生み出す。プロセスの結果として、各I g 重鎖の可変領域はV分節、D分節、及びJ分節を、これらの分節間の接合部に広がる接合配列と共に含む（図4Aに示されるように）。同様に、各I g 軽鎖高頻度可変性領域はV分節及びJ分節並びにこれらの分節間の接合部に広がる接合配列を含む（図4Bに示されるように）。可変領域内で、C D R 1 及びC D R 2 はV分節に見られ、C D R 3 はV分節の幾らか、D分節の全て（重鎖において）、及びJ分節の幾らかを含む。

【0051】

T細胞レセプタは、T細胞の表面に発現する、膜アンカー (membrane anchored) タンパク質である。T細胞レセプタは、同種抗原を認識する結合部分を一緒になって形成する一対のタンパク質鎖を含む。これらは、哺乳類ではC D 3 γ 鎖、C D 3 δ 鎖、及び2つ

10

20

30

40

50

のCD3ε鎖を含む定常T細胞コレセプタ鎖CD3との複合体で発現する。定常鎖は、T細胞レセプタ及び定常ζ鎖と関連付けられて、TCR複合体を形成し、T細胞レセプタに抗原結合されたとき、一緒になってシグナルを生成することが可能である。TCRは、2つの高可変鎖、即ちα及びβ鎖（T細胞の大部分における）又は代替的にはγ及びδ鎖（T細胞の小部分における）を含むヘテロ二量体タンパク質である。各鎖は、2つの細胞外ドメイン：可変領域（又は可変ドメイン）及び定常領域（又は定常ドメイン、細胞膜に近い）と、膜貫通領域と、短細胞質尾部と含む。αβ TCRの場合、MHC（主要組織適合性複合体）分子の状況内で、可変領域は一緒になってペプチド（抗原）に結合する。各可変ドメインは、一緒になって抗原結合部位を形成する、相補性決定領域（CDR、各鎖でそれぞれCDR1、CDR2、及びCDR3と呼ばれる）と呼ばれる3つの高頻度可変性領域を含む。TCRは、BCR及び抗体を含む免疫グロブリンスーパーファミリーのメンバーである。先に説明したものと同様のプロセスにおいて、各TCR鎖の可変領域は幾つかの個片—遺伝子分節（小遺伝子）として知られる—にエンコードされる：β及びδ鎖は可変（V）分節、多様性（D）分節、及び接合（J）分節を含み、α及びγ鎖はV分節及びJ分節を含む。V、D、及びJ遺伝子分節の複数のコピーがゲノムに存在し、T細胞前駆細胞は、V（D）J遺伝子再構成と呼ばれるプロセスで1つのV、1つのD、及び1つのJ遺伝子分節（又はα／γ鎖において1つのV及び1つのJ分節）を（略）ランダムに選択し組み合わせることにより、TCR鎖の可変領域を組み立てる。このプロセスは、必要とされる分節間に、ヘアピンループを形成する2本スタンド切断を形成することを含み、ヘアピンループは次いで一緒に接合される。接合プロセスは不正確であり、V分節とJ分節との間（α／γ鎖）又はV分節とDJ分節とD分節とJ分節との間（β／δ鎖）にヌクレオチドの可変追加又は可変削減が生じ、分節間の接合の配列（「接合配列」と呼ばれる）に大きな多様性を生み出す。V（D）J遺伝子再構成プロセスは、TCRの抗原結合領域に新規のアミノ鎖配列を産生し、抗原認識能力の膨大な多様性を生み出す。プロセスの結果として、各β／δ鎖の可変領域はV分節、D分節、及びJ分節を、これらの分節間の接合部に広がる接合配列と共に含む。同様に、各α／γ鎖高頻度可変性領域はV分節及びJ分節並びにこれらの分節間の接合部に広がる接合配列を含む。可変領域内で、CDR1及びCDR2はV分節に見られ、CDR3はV分節の幾らか、D分節の全て（重鎖において）、及びJ分節の幾らかを含む。

【0052】

本明細書で使用される場合、抗原結合タンパク質の「可変鎖」（本明細書では単に「鎖」とも呼ばれる）は、抗原認識に関わる抗原結合タンパク質の鎖又はその鎖の少なくとも可変領域の部分を含む抗原結合タンパク質鎖の一部を指す。可変鎖は、抗原結合タンパク質内の抗原認識性の多様なレパートリーを担当する可変領域を含む。可変鎖は、BCR重若しくは軽鎖、抗体重若しくは軽鎖、TCRα若しくはβ鎖、TCRγ若しくはδ鎖、又はこれらの鎖内の1つ若しくは複数の可変領域の少なくとも一部を含むそのような鎖の任意の部分であり得る。

【0053】

試料に存在するB細胞レセプタレパートリー（又は対応する抗体レパートリー）は、シーケンシング手法を使用して調べることができる。先に説明したように、2つの主要なシーケンシング手法が使用される：つまり、単一B細胞シーケンシング及びバルクB細胞集団のシーケンシング。BCRシグナリング部分及び抗原結合部分の膜貫通ドメインは可変ではないため、これらの技法は、B細胞レパートリーと対応する抗体レパートリーとの間で共通の部分にフォーカスする。したがって、本開示では、BCR配列、BCRレパートリー、BCR重鎖配列、BCR軽鎖配列、及びそれらの任意の部分への言及は、対応する抗体配列、抗体レパートリー、抗体重鎖配列、抗体軽鎖配列、及びそれらの任意の部分と交換可能に使用される。例えば、BCR重鎖可変領域のシーケンシングへの言及は、対応する抗体重鎖可変領域のシーケンシングと等しく、これら2つの用語は交換可能に使用し得る。「抗原結合タンパク質」という用語は、本明細書では、BCRタンパク質、TCRタンパク質、BCRタンパク質の抗原結合部分、抗体、又は元のBCRタンパク質。TC

Rタンパク質、又は抗体の抗原結合性を維持するそれらの任意の部分を目指すのに使用される。なお、個人の血液中で循環する抗体のレパートリーは、同時に試料に存在するB細胞レセプタレパートリーと一致しなくてよい。これは、もはや個人に存在しないB細胞（例えば、死んだため）により産生された抗体が試料中に存在し得るためである。したがって、「対応する抗体レパートリー」という用語は、試料に実際に存在する抗体（タンパク質）のレパートリーではなく、試料に存在するB細胞により発現される抗体のレパートリーを指す。

【0054】

単一B細胞シーケンシングは、重鎖配列と軽鎖配列との間の対応性を維持することができる。これを行うのに2つの主要な手法を使用することができる。第1の手法は、VH及びVLの物理的リンケージである[DeKosky et al., 2016]。第2の手法は細胞バーコーディング（例えば、10x Genomicsにより提供される等）である[King et al., 2021]。物理的リンケージ手法は、細胞バーコーディング手法よりも高いスループットを有するが、全配列の回収がより難しい。逆に、細胞バーコーディングのスループットはより低い、全配列のより容易な回収が可能である。手法が何であれ、単一B細胞シーケンシングは、先に説明したように、スループットの点で（様々な程度で）制限される。単一B細胞シーケンシング技術によっては、回収される配列の長さに関してさらに制限される。その結果、幾つかの単一B細胞シーケンシング法を使用して同定されたBCR／抗体配列は、重鎖及び軽鎖の両方において、単一のCDR領域、例えばCDR3の調査に制限され得る（換言すれば、隣接V及びJ分節は同定し得るが、V分節におけるCDR1及びCDR2の配列を取得するように完全にはシーケンシングされないことがある）。換言すれば、単一B細胞シーケンシング法からのデータセットは、重鎖及び軽鎖の配列が同定される範囲において変動し得る。シーケンシングされる領域内で、V（D）J分節のあらゆる単一塩基をシーケンシング（又は記録）することは実際的ではないこともあり、したがって、そのようなシーケンシング取り組みは、接合配列と、V、D、及びJ遺伝子の同定に十分な情報との取得にフォーカスし得る。その結果、そのような方法は、以下を含む情報を提供し得る：重鎖のV、D、及びJ分節の同定（例えば、V／D／J遺伝子分節識別子の形態で）、重鎖における接合分節の配列、軽鎖のV及びJ分節の同定（例えば、V／J遺伝子識別子の形態で）、並びに軽鎖における接合分節の配列。各分節の同定は、データベースから対応する生殖細胞系並列を回収するのに使用することができる。しかしながら、データが各分節の同定のみを含む場合、参照生殖細胞系配列と比べて特定の鎖に存在し得る任意の変異（例えば、体細胞変異）は捕捉されないことがある。逆に、バルクB細胞集団のシーケンシングは、重鎖配列と軽鎖配列との間のペアリングを維持しないが、重鎖レパートリー及び軽鎖レパートリー内のそれぞれのシーケンシング能力（特に、BCRレパートリーのシーケンシングの深さ）の点での制限がより緩い。バルクB細胞集団のシーケンシングは、B細胞集団の重鎖レパートリー、軽鎖レパートリー又は両方のシーケンシングを含み得る。しかしながら、上述したように、プロセスのバルク性に起因して、軽及び重鎖レパートリーの両方がシーケンシングされる場合であっても、シーケンシングプロセス中、ペアリング情報を維持することは可能ではない。そのようなシーケンシングは、単一細胞Bシーケンシングを用いて取得されるものと同じ程度まばらな情報又は例えば、完全なCDR配列、複数のCDRの配列、完全な可変領域配列、又は完全な可変領域配列及び配列のアイソタイプを特定するのに十分な定常領域を含むより詳細な情報を生成し得る。同様の考慮事項がT細胞レパートリーのシーケンシングにも当てはまる。特に、B細胞レパートリー及び抗体の研究に関連（特に、これらのレパートリーのシーケンシングに関連）して上述したプロセス及び制限の多くはT細胞レパートリーにも当てはまる。

【0055】

本明細書で使用される場合、「可変鎖配列」という用語は、「重鎖配列」、「軽鎖配列」、「α鎖配列」、「β鎖配列」、「γ鎖配列」、及び「δ鎖配列」の用語を包含し、1つ又は複数の遺伝子分節識別子の組合せ及び／又は一方の末端における接合配列から他方の末端における完全鎖配列に及ぶ、B細胞シーケンシング又はT細胞シーケンシング技術

から取得することができる任意の情報を指す。特に、「重鎖配列」及び「軽鎖配列」という用語は、1つ又は複数の遺伝子分節識別子の組合せ及び／又は一方の末端における接合配列から他方の末端における完全鎖配列に及ぶ、B細胞シーケンシング技術から取得することができる任意の情報を指す。さらに、「可変鎖配列」、「重鎖配列」、「軽鎖配列」、「 α 鎖配列」、「 β 鎖配列」、「 γ 鎖配列」、及び「 δ 鎖配列」という用語は、アミノ酸配列又は対応する核酸コード配列を交換可能に指す。同様に、可変鎖ペアリング又は対（重－軽鎖ペアリング又は対等）は、本明細書に定義されるように、各々が1つ又は複数の遺伝子分節識別子の組合せ及び／又は一方の末端における接合配列から他方の末端における完全鎖配列まで及ぶ、重鎖配列と軽鎖配列との組合せ、 α 鎖配列と β 鎖配列との組合せ、又は γ 鎖配列と δ 鎖配列との組合せを指す。

10

【0056】

例えば、治療用抗体等の所望の抗体又は抗原結合タンパク質を提供する状況内で、「抗体」(Ab)という用語は、所望の生物活性を示し、本明細書に記載のように同定された重－軽鎖ペアリング又は本明細書に記載のように同定された重－軽鎖ペアリングから導出された重－軽鎖ペアリング（例えば、さらなる最適化、親和性成熟等により）を含むモノクローナル抗体、ポリクローナル抗体、多特異性抗体（例えば、二重特異性抗体）、及び抗体断片を含む。

【0057】

「試料」は、本明細書で使用される場合、シーケンシング（例えば、全ゲノムシーケンシング、全エクソームシーケンシング、ターゲット／キャプチャシーケンシング、RNAseq等）等により、ゲノム解析に向けてB細胞ゲノム物質（例えば、RNA又はDNA）を取得することができる細胞又は組織試料、生体体液、抽出物（例えば、被験者から取得されたDNA又はRNA抽出物）であり得る。試料は、被験者から取得（例えば、生検）された細胞、組織、又は生体体液試料であり得る。そのような試料は「被験者試料」と呼ぶことができる。特に、試料は、血液試料、リンパ節試料、脾臓試料、腫瘍試料、又はそれら由来の試料（例えば、B細胞精製、T細胞精製、RNA抽出等）であり得る。本明細書で使用される場合、「ゲノム物質」、「ゲノムシーケンシング」等の用語は、文脈が別段のことを示す場合を除き、ゲノムに存在する物質／配列及び試料のトランスクリプトームの両方を包含する。試料は、被験者から新鮮に取得されたものであってもよく又はゲノム解析前に処理及び／又は貯蔵された（例えば、凍結、固定、又は1つ若しくは複数の精製、濃縮、若しくは抽出ステップを受けた）ものであってもよい。試料は細胞又は組織培養試料であり得る。したがって、本明細書に記載の試料は、被験者から取得された生体試料であるか、それとも例えば細胞株から取得された試料からのものであるかを問わず、B細胞又はそこから導出されたゲノム物質を含む任意のタイプの試料を指し得る。試料は好ましくは哺乳類（例えば、哺乳類細胞試料又はネコ、イヌ、ウマ、ロバ、ヒツジ、ブタ、ヤギ、ウシ、マウス、ラット、ウサギ、若しくはモルモット等の哺乳類被験者からの試料等）、好ましくはヒト（例えば、ヒト細胞試料又はヒト被験者からの試料等）からものである。さらに、試料は輸送及び／又は貯蔵され得、採取は配列データ取得（例えば、シーケンシング）の場所から離れた場所で行われ得、及び／又は本明細書に記載の任意のコンピュータ実施方法ステップは、試料採取場所から離れた場所及び／又はゲノムデータ取得（例えば、シーケンシング）の場所から離れた場所で行われ得る（例えば、コンピュータ実施方法ステップは、「クラウド」プロバイダ等のネットワークコンピュータにより実行し得る）。

20

30

40

【0058】

「配列データ」という用語は、特定の配列を有する試料中のゲノム物質（DNA若しくはRNA）又はプロテオーム物質を示す情報を指す。したがって、配列データは、1つ又は複数のヌクレオチド配列及び／又は1つ又は複数のアミノ酸配列を含み得る。そのような情報は、シーケンシング技術、例えば、次世代シーケンシング（NGS）、例えば、全エクソームシーケンシング（WES）、全ゲノムシーケンシング（WGS）、全トランスクリプトームシーケンシング（RNAseq）、又は捕捉されたゲノム遺伝子座のシーケ

50

ンシング（ターゲット若しくはパネルシーケンシング）等を使用して取得し得る。NGS技術が使用される場合、配列データは特定の配列を有するシーケンシングリード数のカウントを含み得る。配列データは、当技術分野で既知の方法（例えば、ボウタイ（Bowtie）（Langmead et al., 2009）等）を使用して参照配列、例えば、参照ゲノムにマッピングし得る。したがって、シーケンシングリードのカウント又は同等の非デジタルシグナルに特定の場所又は遺伝子座を関連付け得る（「場所」は、配列データがマッピングされた参照ゲノム又はトランスクリプトーム内の場所を指す）。さらに、場所は変異を含み得、その場合、シーケンシングリードのカウント又は同等の非デジタルシグナルに、その特定の場所にける可能なバリエーション（「アレル」とも呼ばれる）の各々を関連付け得る。試料中の特定の場所における変異の存在を同定するプロセスは、「バリエーションコール」と呼ばれ、当技術分野で既知の方法を使用して実行することができる（例えば、GATK Haplotype Caller等の汎用NGSバリエーションコーラー（<https://gatk.broadinstitute.org/hc/en-us/articles/360037225632-HaplotypeCaller>）又はIGBLAST等の免疫配列用に特に設計されたツール（<https://www.ncbi.nlm.nih.gov/igblast/>）[Ye et al., 2013]等）。ゲノム配列データは、当技術分野で既知のように、コード領域をin silicoで翻訳（mRNA配列から直接又はゲノム配列中の同定されたコード領域から）することによりアミノ酸配列に変換し得る。

10

【0059】

本明細書で使用される場合、「処置」は、治療前の症状と比べた、治療中の疾患の1つ又は複数の症状の低減、緩和、又は症状がなくなることを指す。「防止」（又は予防）は、疾患の症状の発現を遅らせること又は防止することを指す。防止は絶対的（疾患が生じないような）であってもよく、又は一部の個人若しくは限られた時間にわたってのみ有効であってもよい。

20

【0060】

本明細書に記載の組成物は、医薬的に許容されるキャリア、希釈剤、又は賦形剤をさらに含む医薬組成物であり得る。医薬組成物は任意選択的に、1つ又は複数のさらなる医薬的活性ポリペプチド及び／又は化合物を含み得る。そのような製剤は例えば、静脈内注入に適した形態であり得る。

【0061】

本明細書で使用される場合、「コンピュータシステム」という用語は、上記実施形態によるシステムを具現するため又は方法を実行するためのハードウェア、ソフトウェア、及びデータ記憶装置を含む。例えば、コンピュータシステムは、中央演算処理装置（CPU）、グラフィカル処理ユニット（GPU）、入力手段、出力手段、及びデータストレージを含み得、これらは1つ又は複数の接続された計算デバイスとして具現し得る。好ましくは、コンピュータシステムは、ディスプレイを有し、又は視覚的出力表示（例えば、ビジネスプロセスの設計での）を提供するディスプレイを有する計算デバイスを備える。データストレージは、RAM、ディスクドライブ、又は他のコンピュータ可読媒体を含み得る。コンピュータシステムは、ネットワークにより接続され、そのネットワークを経由して互いと通信可能な複数の計算デバイスを含み得る。コンピュータシステムがクラウドコンピュータからなってもよく又はクラウドコンピュータを含んでもよいことが明確に考えられる。「プロセッサ」という用語は、特にCPU及びGPUを含む任意の処理ユニット又は処理ユニットの組合せを包含する。

30

40

【0062】

本明細書で使用される場合、「コンピュータ可読媒体」という用語は、限定ではなく、コンピュータ又はコンピュータシステムにより直接読み取りアクセスすることができる任意の1つ又は複数の非一時的媒体を含む。媒体は、限定されないが、フロッピーディスク、ハードディスク記憶媒体、及び磁気テープ等の磁気記憶媒体；光ディスク又はCD-ROM等の光学記憶媒体；RAM、ROM、及びフラッシュメモリを含む、メモリ等の電気記憶媒体；並びに磁気／光学記憶媒体等の上記のハイブリッド及び組合せを含むことができる。

50

【0063】

同定可変鎖対

本開示は、可変鎖配列から可変鎖対を同定する方法を提供する。図1を参照することにより例示的な方法を説明する。図1は、B細胞レセプタ又は抗体の重又は軽鎖配列が使用されて、重軽鎖対を同定する一実施形態を示す。換言すれば、図1は、可変鎖配列がBCRからのBCR／抗体重及び軽鎖である一実施形態を示す。しかしながら、図1を参照することにより説明される方法は、TCR α 、 β 、 γ 、又は δ 鎖配列が $\alpha\beta$ （入力鎖対が α 若しくは β 鎖の場合）鎖対又は $\gamma\delta$ （入力鎖対が γ 若しくは δ 鎖の場合）鎖対の同定に使用される実施形態に適用可能である。任意選択的なステップ10において、B細胞ゲノム物質（典型的にはRNAの形態であり、B細胞ゲノム物質が由来する細胞が発現するBCRのRNAエンコードを抽出しシーケンシングすることができる）を含む試料を被験者から取得し得る。同様に、TCR鎖対が同定される実施形態において、T細胞ゲノム物質を含む試料を使用してもよい。任意選択的なステップ12において、試料中のBCRレパートリーは、バルクBCRシーケンシングを使用してシーケンシングし得る。これは、試料中の重鎖BCRレパートリーをシーケンシングすることを含み得る。同様に、試料中のTCRレパートリーは、バルクTCRシーケンシングを使用してシーケンシングし得る。これは、試料中の β 鎖レパートリーをシーケンシングすることを含み得る。ステップ14において、クエリ鎖配列が提供される。図示の実施形態では、クエリ配列は重鎖配列である。他の実施形態では、クエリ鎖配列は軽鎖配列であり得る。クエリ配列を提供することは、ステップ14Aにおいて、ステップ12においてシーケンシングされた重鎖配列の1つをクエリ配列として選択することを含み得る。クエリ配列を提供することは、ステップ14Bにおいて、V遺伝子配列又は識別子、J遺伝子配列又は識別子、及び接合配列を含む配列を提供することを含み得る。例えば、ステップ14Bは、バルクBCRシーケンシングデータセットから、選択された配列について、V遺伝子配列又は識別子、J遺伝子配列又は識別子、及び接合配列を抽出することを含み得る。同様に、ステップは、例えばクエリ β 鎖配列を使用して、TCRペアリングの状況で実行してもよい。任意選択的なステップ16において、深層学習モデルが提供され、深層学習モデルは、入力としてクエリ可変鎖配列をとり、出力として少なくとも1つの対応可変鎖配列を生成するように構成される。図示の実施形態では、クエリ配列は重鎖配列であり、したがって、深層学習モデルは、入力としてクエリ重鎖配列をとり、出力として少なくとも1つの対応軽鎖配列を生成するように構成される。他の実施形態では、クエリ鎖配列は軽鎖配列であり得、したがって、深層学習モデルは、入力としてクエリ軽鎖配列をとり、出力として少なくとも1つの対応重鎖配列を生成するように構成し得る。さらに他の実施形態では、クエリ鎖配列は β 鎖配列（又は α 、 δ 、又は γ 鎖配列）であり得、したがって、深層学習モデルは、入力としてクエリ β 鎖配列（又は α 、 δ 、又は γ 鎖配列）をとり、出力として少なくとも1つの対応 α 鎖配列（又は少なくとも1つの対応 β 、 γ 、又は δ 鎖配列）を生成するように構成し得る。深層学習モデルは、図示の実施形態では既知の重－軽鎖対からのトレーニング重及び軽鎖配列等の既知の可変鎖対からのトレーニング可変鎖配列を使用して前にトレーニングされていてよい。したがって、深層学習モデルを提供することは単に、方法を実行しているプロセッサと関連付けられたメモリ等のコンピュータ可読媒体から、トレーニングされた深層学習モデルを検索すること又はトレーニング済みの深層学習モデルを他の方法で受信することを含み得る。深層学習モデルのトレーニングについてより詳細に以下説明する。代替的には、深層学習モデルは、図示の実施形態では既知の重－軽鎖対からのトレーニング重及び軽鎖配列等の既知の重－軽鎖対からのトレーニング可変鎖配列を使用して、本方法の一環としてトレーニングし得る。

【0064】

ステップ18において、クエリ鎖配列が深層学習モデルに提供される。ステップ18は、所定のエンコード方式を使用してクエリ配列をエンコードする任意選択的なステップ18Aを含み得る。ステップ18は、所定のエンコード方式を使用して深層学習モデルにより出力された各対応配列をデコードする任意選択的なステップ18Bを含み得る。エンコ

10

20

30

40

50

ードに使用されるエンコード方式及びデコード方式は、深層学習モデルのトレーニングに使用されたトレーニング可変鎖配列の（例えば、重及び軽）鎖配列の内容に基づいて前に定義されていてもよい。ステップ 18 は、1 つ又は複数の対応可変鎖配列の中の最高信頼メトリックが関連付けられた、1 つ又は複数の対応可変鎖配列（図示の実施形態では軽鎖配列）の配列を選択する任意選択的なステップ 18 C を含み得、深層学習モデルは、出力として、確率等の信頼メトリックが各々関連付けられた 1 つ又は複数の対応鎖配列を生成するように構成される。ステップ 18 C は、ステップ 18 B の前に実行されてもよく又は後に実行されてもよい。任意選択的なステップ 20 において、少なくとも 1 つの対応鎖配列（図示の実施形態では軽鎖）における 1 つ又は複数の遺伝子配列識別子は、対応する生殖細胞系配列で置換し得る。任意選択的なステップ 22 において、例えばユーザインタフェースを通して、前のステップのいずれか（特にステップ 18 及び／又は 20）の結果をユーザに提供し得る。これらの結果は、例えば、さらに後述するように、治療用抗体の提供に使用し得る。方法は、複数のクエリ配列に対して繰り返し得る。これはステップ 14 ～ 18 を繰り返すことを含み得る。

10

【0065】

深層学習モデルのトレーニングについて、これより、任意選択的なステップ 10' ～ 16' を参照することにより説明する。ステップ 10' において、既知の可変鎖対からのトレーニング可変鎖配列を含むトレーニングデータが提供される。図示の実施形態では、トレーニングデータは、既知の重－軽鎖対からの重及び軽鎖配列を含む。トレーニングデータは、少なくとも 20,000 個のトレーニング鎖対、少なくとも 30,000、少なくとも 40,000、少なくとも 50,000、少なくとも 60,000、少なくとも 70,000、少なくとも 80,000、少なくとも 90,000、少なくとも 100,000、少なくとも 120,000、又は少なくとも 150,000 個のトレーニング鎖対を含み得る。B 細胞レセプタ／抗体に関連する実施形態では、トレーニングデータは、少なくとも 800,000、少なくとも 100,000、少なくとも 120,000、又は少なくとも 150,000 対のトレーニング重及び軽鎖配列を含み得る。トレーニングデータは、図示の実施形態では重及び軽鎖配列である対になっていないトレーニング配列をさらに含み得る。対になっていないトレーニング鎖配列は「トレーニング前データ」と呼ぶことができる。したがって、トレーニングデータは、トレーニングデータそれ自体（対になった鎖配列、特に、図示の実施形態では対になった重及び軽鎖配列を含む）及びトレーニング前データ（対になっていない鎖配列、特に、図示の実施形態では対になっていない重及び軽鎖配列を含む）を含み得る。トレーニングデータは、少なくとも 100,000、少なくとも 200,000、少なくとも 300,000、少なくとも 400,000、少なくとも 500,000、少なくとも 600,000、少なくとも 700,000、少なくとも 800,000、少なくとも 900,000、少なくとも 100 万（又は少なくとも 500 万、1000 万、1500 万、2000 万、2500 万、3000 万、3500 万、又は 4000 万）個の対になっていない、第 1 のタイプ及び／又は対応するタイプのトレーニング鎖配列を含み得る。トレーニングデータは、少なくとも 100 万（又は少なくとも 500 万、1000 万、1500 万、2000 万、2500 万、3000 万、3500 万、又は 4000 万）個の対になっていないトレーニング重鎖配列と、少なくとも 100 万（又は少なくとも 500 万、1000 万、又は 1500 万）個の対になっていないトレーニング軽鎖配列とを含み得る。当業者は理解するように、トレーニングデータ及び／又はトレーニング前データの量は、利用可能な適したデータの量により制限され得、より多くのデータが利用可能になるにつれて変更し得る。さらに、利用可能なデータ量は、特定の使用事例、例えば、第 1 及び対応鎖配列の同定（例えば、より希である $\gamma\delta$ TCR よりも $\alpha\beta$ TCR のほうが、利用可能なデータが多いことがある）、データをフィルタリングするときに使用される基準（ステップ 12' 参照）等に依存し得る。提供される数は、任意のフィルタリングが適用される前及び／又は後にデータに適用し得る。ステップ 12' において、トレーニングデータはフィルタリングされる。例えば、トレーニングデータは、所定の範囲の長さ外の接合配列（例えば、重及び／又は軽鎖における）を含む

20

30

40

50

あらゆる対を除外するようにフィルタリングし得る。別の例として、トレーニングデータは、例えば、データの導出元の細胞のタイプ、有機体、データがナイーブラライブラリからのものであるか否か、データが、特定の抗原の免疫を有する被験者からのものであるか否か等を含め、データの任意の特徴に基づいてフィルタリングし得る。換言すれば、トレーニングデータは、トレーニングデータが対象となる1つ若しくは複数の特徴を有するデータのみを含む（包含フィルタ）か又は当該データを含まない（除外フィルタ）ことを保証するようにフィルタリングし得る。ステップ14'において、トレーニング鎖配列をエンコードするためのボキャブラリ、特に、図示の実施形態では、トレーニング重鎖配列をエンコードするためのボキャブラリ及びトレーニング軽鎖配列をエンコードするためのボキャブラリを取得することにより、1つ又は複数のエンコード方式がトレーニングデータに定義される。エンコード方式を定義することは、トレーニング鎖配列の内容に基づいて構築されたボキャブラリから、トレーニングデータ内での使用回数が所定の閾値（例えば、2）を下回る任意のトークンを除外することを含み得る。ステップ16'において、トレーニングデータは、入力としてクエリ重鎖配列（図示の実施形態では）をとり、出力として少なくとも1つの対応軽鎖配列（図示の実施形態では）を生成するように、トレーニングデータを使用して深層学習モデルをトレーニングするのに使用される。深層学習モデルをトレーニングすることは、まず、対になっていないトレーニング鎖配列を使用してトランスフォーマベースのモデルをトレーニングすることと、予めトレーニングされたトランスフォーマモデルを使用して、深層学習モデルのエンコーダ及びデコーダを初期化することとを含み得る。代替的には、深層学習モデルをトレーニングすることは、第1のタイプ及び第2のタイプのそれぞれ対になっていないトレーニング鎖配列（図示の実施形態では、第1のタイプは重鎖であり、第2のタイプは軽鎖である）を使用して第1及び第2のトランスフォーマベースのモデルをトレーニングすることと、予めトレーニングされたトランスフォーマモデルを使用して、深層学習モデルのエンコーダ及びデコーダをそれぞれ初期化することとを含み得る。深層学習モデルをトレーニングすることは、既知の鎖対からの鎖配列が上記可変領域に全長配列を含まない場合、欠損配列情報を推定することにより、第2のタイプの鎖及び／又は第1のタイプの鎖（例えば、図示の実施形態では軽鎖及び／又は重鎖）の可変領域に全長配列を含むトレーニング配列を取得することを含み得る。

10

20

30

40

50

【0066】 システム

図2は、本開示による、入力可変鎖から可変鎖対を同定するシステムの一実施形態を示す。システムは計算デバイス1を備え、計算デバイス1はプロセッサ101及びコンピュータ可読メモリ102を備える。図示の実施形態では、計算デバイス1はユーザインタフェース103も備え、ユーザインタフェース103は画面として示されているが、例えば、可聴信号又は視覚的信号等を通して情報をユーザに伝達する任意の他の手段を含み得る。計算デバイス1は、例えばネットワーク6等を通して、シーケンシング機等の配列データ取得手段3及び／又は配列データを記憶した1つ又は複数のデータベース2に通信可能に接続される。1つ又は複数のデータベースは、計算デバイス1により使用し得る他のタイプの情報、例えば、参照配列、パラメータ等をさらに記憶し得る。計算デバイスは、スマートフォン、サーバ、タブレット、パーソナルコンピュータ、又は他の計算デバイスであり得る。計算デバイスは、本明細書に記載のように、入力可変鎖（適宜、重鎖又はβ鎖、好ましくは重鎖）から可変鎖対を同定する方法を実施するように構成される。代替の実施形態では、計算デバイス1は、それ自体が、本明細書に記載のように、入力可変鎖から可変鎖対を同定する方法を実施するように構成されたリモート計算デバイス（図示せず）と通信するように構成される。そのような場合、リモート計算デバイスは、方法の結果を計算デバイスに送信するように構成することもできる。計算デバイス1とリモート計算デバイスとの間の通信は、有線又は無線接続を通してであり得、ローカル又はパブリックネットワークを経由して、例えば、パブリックインターネット又はWiFi等を経由して行われ得る。配列データ取得手段3は、計算デバイス1と有線接続されていてもよく又は無線

接続を通して、例えば図示のようにネットワーク 6 等を通して通信可能であってもよい。計算デバイス 1 と配列データ取得手段 3 との間の接続は、直接であってもよく又は間接（例えば、リモートコンピュータを通して等）であってもよい。配列データ取得手段 3 は、核酸試料、例えば、流体及び／又は組織試料（例えば、末梢血、脾臓、リンパ節、腫瘍組織、又は B 細胞若しくは T 細胞を含む任意の他のタイプの試料等）から精製された B 細胞又は T 細胞から抽出されたゲノム DNA 試料又は RNA 試料から配列データを取得しように構成される。幾つかの実施形態では、試料は、DNA／RNA 精製、断片化、ライブラリ準備、標的配列捕捉（例えば、エクソン捕捉及び／又はパネル配列捕捉等）等の 1 つ又は複数の前処理ステップを受けていてもよい。B 細胞レセプタ配列又はレパートリーの特定に使用するのに適した任意の試料準備プロセスが、本発明の状況内で使用可能である。配列データ取得手段は好ましくは、次世代シーケンサである。配列データ取得手段 3 は、配列データ（生又は部分的に処理された）を記憶し得る 1 つ又は複数のデータベース 2 と直接又は間接的に接続し得る。

10

【0067】

用途

上記方法は、重鎖、軽鎖、又はそれらの一部（例えば、V 遺伝子、J 遺伝子、及び接合配列等）に限られた情報から、標的に結合する可能性が高い抗体又は BCR を同定することが望ましい任意の状況で適用される。これは往々にして、抗体治療の発見プロセスの状況で該当する。抗体治療は、神経変性疾患からがんまで広範囲の疾患に対して成功する手法であることがわかっている。したがって、本明細書に記載の手法は、これらの臨床状況の各々で治療を提供する状況で適用される。さらに、本明細書に記載の方法は、入力情報が特定の目的で新たに生成される（例えば、所望の表現型を有すると同定された患者又は試料から）か、それとも既存／過去のデータセットからのものである（例えば、既存のデータセットをマイニング又は再マイニングして、新たな治療を発見又は特定の臨床表現型が何故存続するのかを説明することができる免疫タンパク質を同定するため）かを問わず、任意の入力重／軽鎖又はその一部から潜在的に機能する抗体又は BCR を同定するのに使用することができる。

20

【0068】

したがって、本発明は、抗体治療を提供する方法も提供し、方法は、本明細書に記載の任意の方法を使用して重－軽鎖ペアリングを同定又は本明細書に記載の任意の方法を使用して同定された重－軽鎖ペアリングから導出（例えば、さらなる最適化、変異等により等）された重－軽鎖ペアリングを同定することを含む。重－軽鎖ペアリングは、1 つ又は複数の試料中の重鎖レパートリーのバルク BCR シーケンシングにより取得された入力重鎖配列に対して取得し得る。1 つ又は複数の試料は、1 人又は複数人の被験者からの試料であり得る。1 人又は複数人の被験者は、所望の特性、例えば、特定の臨床表現型又はバイオマーカープロファイル等の臨床的に関連する特性等を有すると同定されていることがある。例えば、1 人又は複数人の被験者は、特定の疾患又は症状に対する回復力を有し得る。疾患又は症状は、がん（例えば、乳がん等）、神経変性疾患（例えば、筋萎縮性側索硬化症等）、又は感染疾患（例えば、COVID-19 等）から選択し得る。方法は、1 つ又は複数の試料において同定された重鎖配列から選択された複数の入力重鎖配列の重－軽鎖ペアリングを同定し、それにより、重－軽鎖ペアリングのセットを取得することを含み得る。方法は、重－軽鎖ペアリング又は重－軽鎖ペアリングのセット内の各重－軽鎖ペアリングの標的（又は複数の推定上の標的又は標的のセット）を同定することをさらに含み得る。方法は、複数のペプチド候補と突き合わせて、1 つ又は複数の試料と同じソースからの抗体をスクリーニングすることにより、1 つ又は複数の標的を同定することをさらに含み得る。複数のペプチド候補は、1 つ又は複数の試料の由来元である種に基づいて選択し得る。例えば、1 つ又は複数の試料のソースは、1 人又は複数人のヒト被験者であり得、1 つ又は複数の試料と同じソースからの抗体レパートリーは、ヒトペプチドームを表すペプチド候補のセットと突き合わせてスクリーニングし得る。重－軽鎖ペアリング又は重－軽鎖ペアリングのすることと内の各重－軽鎖ペアリングの標的（又は推定上の標的又は標

30

40

50

的のセット)を同定することは、複数のペプチド候補と突き合わせて、1つ又は複数の試料と同じソースからの抗体をスクリーニングすることにより同定された1つ又は複数の標的を使用することを含み得る。方法は、1つ又は複数の基準に基づいて重-軽鎖ペアリングのセットをフィルタリングすることをさらに含み得る。1つ又は複数の基準は、推定上の標的又は重-軽鎖ペアリングに同定された法的のセットの同定に適用し得る。方法は、同定された重-軽鎖ペアリング又は同定された重-軽鎖ペアリングから導出された重-軽鎖ペアリングを含む抗体又はその断片を取得することをさらに含み得る。抗体又はその断片を取得することは、抗体又はその断片のコード配列を同定することと、その配列を適した発現系(例えば、適した宿主細胞等)において発現させることとを含み得る。方法は、例えば、1つ又は複数の抗原候補への結合について検定することにより、抗体又はその断片が結合する1つ又は複数の抗原を同定することをさらに含み得る。方法は、抗体又はその断片の配列を最適化することをさらに含み得る。抗体又はその断片の配列を最適化することは、当技術分野で既知の任意の抗体最適化技法を使用して実行し得る。抗体又はその断片の配列を最適化することは、例えば、重-軽鎖ペアリングが同定された入力配列と同様に配列を解析することにより重-軽鎖ペアリングが同定された配列データからの情報を使用して実行し得る。

10

【0069】

本発明は、免疫治療組成物を提供する方法も提供し、方法は、本明細書に記載のように重-軽鎖ペアリングを同定することと、重-軽鎖ペアリングを含む抗体又は重-軽鎖ペアリングから導出された抗体(例えば、さらなる最適化、変異等により)を含む免疫治療組成物を生成することを含む。

20

【0070】

本明細書に記載の方法は、二重特異性抗体を提供する状況で利用することもできる。例えば、本明細書に記載の方法は、対象となる2つの異なる重鎖とのペアリングに適した軽鎖を同定するのに使用し得る。したがって、本発明は二重特異性抗体を提供する方法も提供し、方法は、本明細書に記載の任意の方法を使用して2つの重鎖の各々に共通する軽鎖ペアリング又は本明細書に記載の任意の方法を使用して同定された重-軽鎖ペアリングから導出された(例えば、さらなる最適化、変異等により)共通する軽鎖と2つの重鎖との組合せを同定することを含む。そのような実施形態では、深層学習モデルが、複数の対応する軽又は重鎖配列を出力することが有利であり得る。例えば、深層学習モデルを使用して、第1の重鎖配列に対する第1の複数の対応する軽鎖配列を予測し、第2の重鎖配列に対する第2の複数の対応する軽鎖配列を予測し得る。予測された第1及び第2の複数の軽鎖配列は次いで比較されて、両方の重鎖を含む二重特異性抗体の共通する軽鎖として使用するのに適し得る1つ又は複数の軽鎖を同定し得る。

30

【0071】

本明細書に記載の方法は、抗体最適化の状況で利用することもできる。例えば、本明細書に記載の方法は、重鎖とペアリングするのに適する軽鎖を同定するのに使用し得、ペアリングは、重鎖の元のペアリングと比べて1つ又は複数の有利な性質(例えば、機能性又は開発可能性の改善等)を有する。したがって、本発明は、改良された抗体を提供する方法も提供し、方法は、元の抗体の入力重鎖から本明細書に記載の任意の方法を使用して重-軽鎖ペアリングを同定又は本明細書に記載の任意の方法を使用して同定された重-軽鎖ペアリングから導出される(例えば、さらなる最適化、変異等により)重-軽鎖ペアリングを同定することを含む。本明細書に記載の方法は、 β 鎖、 α 鎖(若しくはあまり一般的ではない場合は、 γ 若しくは δ 鎖)、又はそれらの一部(例えば、V遺伝子配列、J遺伝子配列、及び接合配列等)に限られた情報から、標的に結合する可能性が高いTCRを同定することが望ましい任意の状況でも利用される。これは往々にして、遺伝子操作されたT細胞等の細胞治療の発見プロセスの状況で当てはまる。したがって、本発明は、特定のTCRを発現する遺伝子操作されたT細胞等のTCRベースの治療を提供する方法も提供し、方法は、本明細書に記載の任意の方法を使用して $\alpha\beta$ 若しくは $\gamma\delta$ 鎖ペアリングを同定又は本明細書に記載の任意の方法を使用して同定された $\alpha\beta$ 若しくは $\gamma\delta$ 鎖ペアリン

40

50

グから導出（例えば、さらなる最適化、変異等により）された $\alpha\beta$ 若しくは $\gamma\delta$ 鎖ペアリングを同定することを含む。したがって、本明細書に記載の方法は、抗体について上述したのと同様に、T細胞レセプタ最適化の状況で利用することもできる。

【0072】

以下は例として提示され、特許請求の範囲の限定として解釈されるべきではない。

【実施例】

【0073】

実施例

これらの例は、本発明により重-軽鎖ペアリングを同定する方法を説明し、既知のペアリング及び新たに取得された実験データを有する単一細胞データセットを使用して検証する。

10

【0074】

方法

データセット

全てのデータセットの概要統計情報が以下の表1に記載される。トレーニング、検証、検定、及びブラインドテストセットのソース及び階層を以下説明する。

【0075】

【表1】

表1. 研究に記載されるデータセットの概要。

20

データセット	方法論	ドナー	一意の重-軽鎖対の数	参照
トレーニング	単一細胞エマルジョン	6	153889	DeKosky ら, 2015 年; DeKosky ら, 2016 年
検証	単一細胞エマルジョン	6	17099	DeKosky ら, 2015 年; DeKosky ら, 2016 年
検定	単一細胞エマルジョン	6	18999	DeKosky ら, 2015 年; DeKosky ら, 2016 年
Eccles	10x 単一細胞	1	741	Eccles ら, 2020 年
King	10x 単一細胞	7	30332	King ら, 2021 年
Setliff	10x 単一細胞	2	4944	Setliff ら, 2019 年

30

【0076】

トレーニングセット、検証セット、及びテストセット：DeKosky et al., [2015]において3人のドナーからの対になった重-軽鎖配列と、DeKosky et al. [2016]からの3つのナイーブBCRライブラリとを組み合わせた。これらのデータセットは、各々が以下を含むエントリを含む：重鎖V遺伝子識別子、重鎖接合配列（ヌクレオチド及びアミノ酸）、重鎖J遺伝子識別子、軽鎖V遺伝子識別子、軽鎖接合配列（ヌクレオチド及びアミノ酸）、及び軽鎖J遺伝子識別子。このトレーニングセットは主に、公衆での利用可能性及びサイズに起因して選ばれた。なお、データエントリは、重鎖D遺伝子識別子も含んだが、この情報は使用されなかった。これは、D遺伝子のアノテーションが、V遺伝子及びJ遺伝子のアノテーションよりも正確性が低いと考えられているためである。配列をフィルタリングして、長さがアミノ酸7～30個の重鎖接合部及び長さがアミノ酸7～20個の軽鎖接合部を有するものを残した。これは、これらの境界外の接合配列は希であると考えられており、長さが増えると、希であるより長い配列に対する情報における利得でバ

40

50

ランスがとられる可能性が低い計算力に関するコストを伴うためである。このフィルタは、境界外のサイズを有する重鎖接合部を有した 84 対及び境界外のサイズを有する軽鎖接合部を有する 168 対を除去した。より簡潔なボキャブラリを維持するために、重鎖 V 遺伝子及び軽鎖 V 遺伝子が少なくとも 2 つの配列で観測された配列についてデータをフィルタリングした。実際に、ボキャブラリサイズの増大は、非常に希な V 遺伝子を含むことにおける恩益でバランスがとられる可能性が低い、追加のモデルパラメータをトレーニングするコストを伴う。このフィルタは、少数のエントリしか除去しなかった。実際に、開始対の中から 1 つの重鎖 V 遺伝子のみが 1 度観測された。全ての軽鎖は少なくとも 2 つの配列で観測された。IMGT 偽遺伝子を有する配列（コード領域が終止コドン及び／又はフレームシフト変異及び／又は開始コドンに影響する変異を有する配列である、免疫遺伝学データベース（<http://www.imgt.org>）において偽遺伝子としてラベルされた配列を含むエントリ）も除去した。最後に、任意の全く同じ重－軽鎖対を除去し、これは、99%冗長性カットオフと機能的に同等である（換言すれば、異なる複数のセットにわたる任意の重－軽鎖対が最大限で 99% 同一であり、トレーニングデータセット内の対が実際に一意であることを確認する）。フィルタと一緒に除去したエントリの長さ及び数は、偽遺伝子フィルタ及び冗長性フィルタを通過した 190, 240 の開始対からの合計 253 対（84 + 168 + 1）、189, 987 対のセットが残った（表 1 参照）。

10

【0077】

189987 個の配列を 153889、17099、及び 18999 個の配列のトレーニングセット、検証セット、及びテストセットに分けた（約 80% トレーニング／10% 検証／10% 検定の分割に対応する）。同じ重鎖配列が 3 つのセットにわたって存在することができるが、これらの重鎖のいずれも同一の軽鎖相手方を持たない。

20

【0078】

トレーニングデータ調査：トレーニングセット内に少なくとも 5 つの観測がある重－軽鎖対に対して χ^2 検定を使用してトレーニングデータの特徴を調べた。これは、重－軽鎖ペアリングがランダムではない（ $\chi^2 = 8102.9$ 、 p 値 $< 10^{-9}$ ）ことを示した。しかしながら、重－軽鎖対の分割表は研究間で一貫せず、ペアリングのランダム性に関する推論の結論は出なかった。配列レベルにおいて、単一の重鎖配列は典型的には、1 つの一意的相手方軽鎖配列を有したが、1 重鎖につき最高で 7 つの異なる軽鎖相手方が存在することができる。軽鎖はより雑多であることがわかり、単一の軽鎖配列は最高で 1042 個の異なる重鎖と対になることがわかった。それにもかかわらず、73549 個のうち 58123 個の軽鎖が 1 つの一意的重鎖相手方を有した。

30

【0079】

ブラインドテストセット：単一細胞データセットを対になった (Paired) Observed Antibody Space から取得した [Kovaltsuk et al., 2018]。King et al., [2021]、Eccles et al., [2020]、及び Setliff et al., [2019] からのヒト BCR レポートリーの 3 つの研究を集めた。それぞれ合計で 30332、741、及び 4944 個の一意的重－軽鎖対が存在した。

【0080】

配列トークン化

40

トレーニングセットデータは、全長重鎖及び軽鎖配列を含まず、その理由は、このデータの生成に使用された単一細胞シーケンシング法が軽／重鎖の完全なアミノ酸配列を回収することができなかったためであった。その代わりに、トレーニングデータセットは以下を含んだ：(i) 各重鎖について：V 遺伝子識別子、接合配列、J 遺伝子識別子、及び D 遺伝子識別子（しかし D 遺伝子識別子は使用されなかった）及び (ii) 各軽鎖について：V 遺伝子識別子、接合配列、及び J 遺伝子識別子。したがって、各エントリは、例えば、IGHV3-23/CAR...DYW/IGHJ6-IGKV3-20/CQQ.../IGKJ2 等の遺伝子識別子と配列との組合せを含んだ。モデルは、V 遺伝子識別子、接合配列、及び J 遺伝子識別子に対応するトークン化された重鎖配列を入力としてとり、出力として、V 遺伝子識別子、接合配列、及び J 遺伝子識別子に対応するトークン化され

50

た軽鎖配列を生成するようにトレーニングされた。このデータ形式に対処するために、トークン化用にカスタムエンコード法を設計した。各V遺伝子は単一のトークンからなり、各J遺伝子は単一のトークンからなり、接合アミノ酸配列は重複3-merとしてトークン化された。接合配列は、配列の最も多様性に富む領域であり、結合機能の大半を仲介すると考えられており、したがって、この配列のトークン化の粒度を上げた。トレーニングセットで少なくとも2回発生した場合、トークンを使用した（データのフィルタリングに関連して既に上述したように）。

【0081】

接合アミノ酸配列（又は完全な配列）のトークン化に可能な他の方式には、例えば、バイト対エンコード又は各アミノ酸のトークンがある。バイト対エンコード等の方式は、単一細胞データ等のより多くの完全配列の配列データが利用可能な場合、例えば、数十万のオーダ又は100万もの配列がモデルのトレーニングに使用される場合等、特に有用であり得る。

【0082】

合計で、7986個のトークンを重鎖ボキャブラリ内で構築し、一方、軽鎖ボキャブラリは6452個のトークンを有した。この例では、「センテンス」は、重又は軽鎖配列のトークン化された表現である。各センテンスは特殊なトークン<SOS>で始まり、その後、重又は軽鎖のV遺伝子を表すトークン、重複3-merトークン、J遺伝子トークン、そして次に特殊なトークン<EOS>が続く。重鎖及び軽鎖のそれぞれの最大長34及び24よりも少ないトークンを有するあらゆるセンテンスでは、配列を特殊な<PAD>トークンでパディングした。配列が未観測のV遺伝子、J遺伝子、新規の3-merを含む場合又は接合部がアミノ酸30個よりも長い場合、Matchmakerは予測を行わない。実際に、J遺伝子多様性はV遺伝子多様性よりもはるかに低く、したがって、未観測J遺伝子のインスタンスは存在しなかった。

【0083】

軽鎖ペアリング方法

入力として重鎖のみを使用した軽鎖予測の5つの戦略を調べた。これらは図3に示されており、さらに以下詳述する。戦略は以下を含む：

- － 配列への重鎖トークン化及び軽鎖トークン化変換を用いるトランスフォーマーアーキテクチャ（図3A及び以下において「Matchmaker」と呼ばれる）；
- － トランスフォーマーと同じボキャブラリを使用し、これもまた軽鎖出力（図3B）を提供する、重鎖入力を用いるGRU（ゲート付き回帰型ユニット）モデル；GRU及びより広くはアテンションを用いるLSTM（長短期記憶ネットワーク）等の他の回帰型ニューラルネットワーク（RNN）アーキテクチャは、トランスフォーマーがこのタスクに一般的になる前、ニューラルマシン翻訳の前の「現行技術水準」であった；
- － データベース探索法ワークフロー（図3C）；
- － 「ランダム探索」と呼ばれるデータベース探索法の変形；及び
- － 重及び軽鎖リードカウントのランク付き分布をマッチングし、次いで同様のランクを有する鎖をペアリングする頻度探索法（図3D）。

【0084】

これらの方法のうちの4つ（トランスフォーマー、GRU、データベース探索、ランダム探索）は、この研究の一環として新たに開発された。頻度探索手法は、Reddy et al., 2010において前に記載された手法に基づく。

【0085】

Matchmakerアーキテクチャ及び推論：MatchmakerはPyTorch（バージョン1.6.0）を用いて構築された。Matchmakerのハイパーパラメータ及び最適化手順は、Vaswani et al. [2017]からの配列-配列（Seq2Seq）トランスフォーマーと同様である。Vaswani et al. [2017]のオリジナルのトランスフォーマーからの違いについて以下説明する。まとめると、モデルは、よりよい制約付けのためにわずかに小さく（数層）作られ、トレーニングを改善するために異なる最適化技法を使用してトレーニングされ

た。Matchmakerは、4つのエンコーダ層及び4つのデコーダ層を有し、フィードフォワード次元1024及びドロップアウト0.2を有する。層正規化は残差ブロック内で適用された[Child et al., 2018; Xiong et al., 2020]。モデルはAdamWを使用して最適化され、重み減衰は0.1を用いた。1.0のL2ノルムを用いて勾配クリッピングを実施した。Matchmakerは、合計で31.7M個の学習可能なパラメータを有する。トレーニングは、検証損失が3エポックにわたり改善しなかった場合に停止し、最良の検証損失を有するモデルを使用した。

【0086】

ゲート付き回帰型ユニット (Gated recurrent unit) (GRU) ニューラルネットワークモデル：アテンションメカニズムを用いた代替深層学習Seq2Seqモデル [Bahdanau et al., 2015] は、2つのGRUネットワークを使用してトレーニングされた [Cho et al., 2014]。ここで、エンコーダは隠れ次元1024を有する4層双方向GRUであり、デコーダは隠れ次元1024を有する4層の順方向のみGRUである。埋め込み層の次元等のモデルの他のハイパーパラメータは、可能な限りMatchmakerと近くなるようにマッチングした。合計で、このモデルは131.8M個の学習可能なパラメータを有する。エンコーダーデコーダGRUモデルは、Matchmakerと同一の様式でトレーニングされた。簡潔にするために、このアーキテクチャを「GRUモデル」と呼ぶ。

【0087】

GRUネットワーク (及び拡大解釈により回帰型ニューラルネットワーク) は、トランスフォーマと比べて、配列をいかに処理するかメカニズムが全く異なる。手短に言えば、トランスフォーマは、より高速にするのみならずより正確にする一連の「自己アテンション」メカニズムを使用し、一方、回帰型ネットはこれを全く持たない。

【0088】

この研究で使用されたMatchmakerモデル及びGRUモデルは両方とも、各入力鎖に単一の予測を提供した。特に、モデルは、続く各位置において、現在位置で可能な全てのトークンの確率を出力し、現在位置で最高確率を有するトークンを選択してから、次の位置に移ることにより、各鎖を逐次予測した。他の実施態様も可能であり考えられる。例えば、複数のトークンを各位置で同時に考慮することができ、複数の位置 (例えば、配列全体、即ち全ての位置等) にわたる大域的確率を最適化する複数の位置にわたるトークンの組合せを選択することができる。これは、ビームサーチ手法又はビームスタックサーチ等の関連する手法 [Zhou & Hansen, 2005] 及び深さ優先ビームサーチ [Furcy & Konig, 2005] を使用して実行し得る。ビームサーチは、限られたセット内の最も見込みのあるノードを拡大することによりグラフを探索するヒューリスティック探索アルゴリズムである。所定の最大深さに達した後、最大確率を有する解を出力し得る。データベース探索法：Matchmakerのトレーニングセット内の既知の重-軽鎖対に相同する配列により重鎖をペアリングした。クエリ重鎖配列の場合、クエリに対して一致する重鎖V遺伝子を有する重-軽鎖対のみが選択される。このサブセットから、2対が選択される。第1の対は、クエリに対して最も近い重鎖接合アミノ酸配列を有する対からのものである「最も近い」軽鎖である。第2の対は、クエリ重鎖のV遺伝子と関連付けられる頻度が最も高い軽鎖V遺伝子を有する対からのものである「上位」軽鎖である。クエリ接合アミノ酸配列への同定が65%以上である場合又は最も近い軽鎖と上位軽鎖との間の生殖細胞系V遺伝子同定が75%以上の場合、最も近い軽鎖が使用される。その他の場合、上位軽鎖が使用される。これら2つの事例を区別する論拠は、探索において十分に類似する接合配列が存在する場合、これからのVL配列を使用することができることである。十分に類似するものがない場合、そのV遺伝子に最も一般的なVLが使用されるより粗い手法がとられる。異なるカットオフを使用してもよく又はカットオフを使用しないことさえ可能である (即ち、全事例で最も近い軽鎖を使用する)。いずれの場合でも、この戦略は「データベース探索」法と呼ばれる。

【0089】

ランダム探索法：この手法を用いる場合、重鎖の特徴を全く考慮せずに、1つの軽鎖が

データベースからランダムに選ばれる。

【0090】

頻度ベースの探索法：ベースラインとして、軽鎖対の探索はリードカウントを使用して実行された。この研究で使用された全てのデータセットは予め対になっているため、2つのバルクシーケンシングライブラリがある状況は、まず、対になった配列をばらばらにすることによりエミュレートされた。次いで、リード数を重鎖配列又は軽鎖配列上に集約した。例えば、4つのリードを有する配列Heavy A: Light A及び5つのリードを有する配列Heavy B: Light Aがあると考え、分割及び集約により、4つのリードを有するHeavy A、5つのリードを有するHeavy B、及び9つのリードを有するLight Aが生成される。重鎖及び軽鎖は次いで、合計リードカウントに基づいてランク付けられる（図3D参照）。重鎖は次いで、ランクが一致する軽鎖とペアリングされる。

10

【0091】

完全軽鎖配列の構築：上述した全ての方法で、出力は軽鎖V遺伝子識別子、接合アミノ酸配列、及び軽鎖J遺伝子識別子である（図3）。トレーニングセットが全長配列を持たない場合、軽鎖V遺伝子及びJ遺伝子をそれらの生殖細胞系アミノ酸配列で置換した。

【0092】

同種標的抗原へのモノクローナル抗体結合の特徴付け

各治療用抗体の計算により対になったものは、間接的ELISAにより各同種抗原への結合について特徴付けられた。手短に言えば、抗原の商業的なソースからの組換え版（以下の表2参照）を固定化し、最高濃度500 nM及び3倍希釈系列を使用した10点滴定として抗体を溶液中に加えた。HRP（西洋ワサビペルオキシダーゼ）に共役した抗ヒトIgG二次抗体を使用して、洗浄後もなお結合していた抗体を検出した。TMB-ELISA基質を用いて現像し、硫酸を用いて停止した後、450 nm吸光度を読み取った。次いで、GraphPad Prism 9（www.graphpad.com）を使用して濃度と突き合わせて吸光度をプロットした。

20

【0093】

【表2】

表2. ELISAに使用した、商業的ソースからの抗原

30

抗原	コンストラクト	カタログ番号
C-term His タグを有する遺伝子組換えヒト ErbB2/Her2 Fc キメラ	Thr23-Thr652 Acc# NP_004439	1129-ER
遺伝子組換えヒト PD-L1/B7-H1 Fc キメラ	Ph19-Thr239 Acc# Q9NZQ7	156-B7
遺伝子組換えヒト PD-1 Fc キメラ	Leu 25-Gln167 Acc# Q15116	1086-PD
遺伝子組換えヒト EGFR Fc キメラ	Leu25-Ser645 Acc#CAA25240	344-ER
遺伝子組換えヒト Siglec-3/CD33 Fc キメラタンパク質	Asp18-His259 Acc# AAA51948	1137-SL

40

【0094】

モノクローナル抗体の熱安定性

モノクローナル抗体の熱安定性を熱変性アッセイで測定した。トリプリケートにおいて、SYPRO（商標）オレンジの存在下で各抗体を25℃から95℃まで加熱し、蛍光度を測定した。次いで、3つの反復の平均として溶融曲線導関数をプロットした。蛍光度が最も急に増大した温度を各抗体で記した。

【0095】

タンデムトランスフォーマモデル

上述した5つの軽鎖ペアリング手法に加えて、さらなるトレーニングデータを利用する

50

ことが可能な別の手法を設計した。特に、（全長）B C R重及び軽鎖配列（概ね4 2 3 0万個の重鎖配列及び1 5 3 0万個の軽鎖配列）でそれぞれ予めトレーニングされた2つのR o B E R T aモデル [Liu et al., 2019; Rothe et al., 2020] を含むタンデムトランスフォーマモデルを構築した。Matchmaker及びG R Uモデルと異なり、このモデルは、トークンとして各アミノ酸を使用してトレーニングされ、したがって、配列の完全なカバレッジを可能にした。対になっていない抗体配列でトレーニングされた、この予めトレーニングされたR o B E R T aモデルは、本明細書では「A n t i B E R T a」と呼ばれ、トレーニング手順を図9に記載する。A n t i B E R T aモデルは、R o B E R T aベース (Liu et al., 2019) と同様のスタイルでトレーニングされたが、より小さなバッチサイズ7 6 8、ピーク学習速度 10^{-4} 、並びに2 2 5 0 0 0を超えるトレーニング前ステップ及び1 0 0 0 0のウォームアップステップを用いた。予測では、ビームサーチ (Sustkever et al., 2014) をビーム幅3で使用した。

10

【0096】

2つのA n t i B E R T aモデルを配列-配列モデルとして結合し、エンコーダ及びデコーダは各々、Huggingfaceトランスフォーマライブラリ [Wolf et al., 2019] を使用してA n t i B E R T aモデルのコピーとして初期化された。A n t i B E R T a-A n t i B E R T aモデルを次いで、上述したMatchMaker及びG R Uモデルで使したものよりもわずかに大きな、対になった配列のデータセットを用いて微調整した。特に、トレーニングデータは、抗原経験ライブラリ（先に説明したものと同一ソースから、即ち、DeKosky et al. 2016及びDeKosky et al., 2015）からより多くの配列を導入することにより拡大した。合計で、1 7 1 9 8 4個の対になった配列が存在した。しかしながら、前と同様に、ここでデータセットは完全な重鎖配列及び完全な軽鎖配列を含まない。したがって、生殖細胞系V及びJ遺伝子アノテーションを使用して完全な配列を推測した。微調整ステップでは、モデルを20エポックにわたり、ピーク学習速度 3×10^{-5} 及び5%ウォームアップを用いてトレーニングした。パラメータは、エンコーダA n t i B E R T aとデコーダA n t i B E R T aとの間で共有された。

20

【0097】

予めトレーニングされたいわゆる「チェックポイント」の使用は、N L Pの状況で強力な戦略であることがわかっている [Rothe et al., 2020]。特に、B E R T-B E R Tアーキテクチャは、Rothe et al. [2020]においてNMTに対して上手く機能することが示されていた。トランスフォーマモデルを使用してNMT問題として重-軽鎖ペアリングの問題に取り組むこと（以下参照）を示してきたが、本発明者らは、B E R T-B E R Tアーキテクチャも非常に上手く機能する可能性が高く、R o B E R T aモデルも種々のN L PタスクでB E R Tモデルを改善することが示されたため、R o B E R T a-R o B E R T aアーキテクチャがこれをさらに改善し得るという仮説をたてた。

30

【0098】

フルモデル（予めトレーニングされたA n t i B E R T aモデルを含む）のトレーニングに使用される対になったトレーニングデータは、全長B C R重鎖配列及び軽鎖配列を含まないため（上述したように）、全長配列を含む同等の対になったデータセットが、2つの代替の手法のうちの一方を使用して生成される。第1の手法では、全長配列は、V及びJ遺伝子識別子に対応する生殖細胞系配列で置換することにより取得される。第2の手法では、予めトレーニングされたA n t i B E R T aモデル（又は任意の他のそのような「チェックポイント」モデル、例えば、G P T-2モデル等）を使用して、既知の鎖対に基づいて重鎖及び軽鎖で独立して（各モデルを使用して）トレーニングセットの全長配列を予測する。「チェックポイント」モデルからの予測は、最適には完全な配列が利用可能ではない任意の分節の生殖細胞系配列から取得された幾らかの情報（例えば、分節のアミノ酸の幾つか、例えば分節の最初のk個のアミノ酸の同定等であり、kは例えば1、2、3、5、10等であることができる）と組み合わせて、既知の鎖対の一部又は全て（例えば、遺伝子分節識別子、部分配列等）を使用して取得し得る。換言すれば、対になっていない全長重及び軽鎖データでトレーニングされたA n t i B E R T aモデルを使用して、デ

40

50

ータで提供されたV遺伝子識別子、J遺伝子識別子、及び接合配列から、トレーニングデータ内の重鎖の全長配列を予測し得る。同様に、対になっていない全長重及び軽鎖データでトレーニングされたAntibERTaモデルを使用して、データで提供されたV遺伝子識別子、J遺伝子識別子、及び接合配列から、トレーニングデータ内の軽鎖の全長配列を予測し得る。同じ2つの手法を使用して、任意の限られた、対になったトレーニングデータを、「チェックアウト」モデルのトレーニングに利用可能であった可能性があるより拡張された形式のデータにマッピングすることができる。代替的には、「チェックポイント」モデルのトレーニングに使用されるデータは、対になったトレーニングデータの形式に一致する限られた形式に変換し得る。これはなお、利用可能な膨大な数の対になっていない配列から、予めトレーニングされたモデルにより収集された潜在的な追加情報から恩恵を受け得る。しかしながら、そのような対になっていない配列データで利用可能な情報の広がりを完全には利用しないことがある。

10

【0099】

結果

NLPに触発されたモデルは、入力として重鎖のみを使用して軽鎖を生成する

重-軽鎖ペアリングの問題は、NMTタスクとして骨組みが組まれ、入力として重鎖配列のみが与えられて軽鎖配列が予測される。特に、[Vaswani et al., 2017]に類似するSeq2Seqトランスフォーマが実施された。モデルは、方法(図3A)に記載のように、入力として重鎖配列のトークン化表現でトレーニングされ、出力として軽鎖配列のトークン化表現を返す。入力重鎖配列について、モデルエンコーダ層は、V遺伝子、重複接合k-mer、及びJ遺伝子間の自己アテンションを計算する。重鎖配列でのエンコーダの自己アテンションスコアが次いでデコーダにより使用されて、軽鎖配列を自己回帰により予測する。デコーダアテンションの一例が図5に示されており、図5では、各アテンションヘッドは重鎖トークンの異なるサブセットにフォーカスして、出力軽鎖トークンを決定する。

20

【0100】

図6は、軽鎖V遺伝子予測に関して、全ての方法でのヘルドアウトテストセット及び単一細胞ブラインドテストでの予測性能を示す。V遺伝子は、抗原結合を決める鎖の部分の最大部分をなし、したがって、安定性に大きな影響力を持つとともに、結合にある程度影響力を持つため(結合は接合配列によっても大きく影響される)、V遺伝子結果は完全予測結果とは別々に調べられる。軽鎖V遺伝子予測結果についてさらに以下考察する。V遺伝子、接合アミノ酸配列、及びJ遺伝子からなる正しい完全な軽鎖の数は、全ての方法にわたり、正しいV遺伝子数よりもはるかに少なかった。予測結果の完全なセットは表3及び表4にある。

30

【0101】

【表3】

表3. 正しい軽鎖V遺伝子予測数

	Matchmaker	データベース	GRU	頻度	ランダム	合計
検定	2431	4940	1786	658	915	18803
Eccles	64	55	53	26	38	730
King	2671	2345	2119	686	1224	29785
Setliff	470	323	434	143	220	4797

40

【0102】

50

【表 4】

表4. 正しい軽鎖V遺伝子+接合アミノ酸配列+J遺伝子予測数

	Matchmaker	データベース	GRU	頻度	ランダム	合計
検定	139	1162	44	12	11	18803
Eccles	7	4	2	1	2	730
King	105	59	44	0	1	29785
Setliff	23	6	44	0	3	4797

10

【0103】

テストセット内の18999個の配列のうち、18803個は5つ全ての予測方法により予測された。トランスフォーマベースのモデルは、2431個の重鎖配列で正しい軽鎖V遺伝子を予測し（12.9%）、一方、GRUモデルは1786個の正しい予測を有した（9.5%）。トランスフォーマにおける自己アテンションメカニズムの使用は、トランスフォーマモデルが、GRUモデルと比べて、重鎖配列についての情報をよりよく利用可能であるはずであることを意味する。これは、GRUモデルと比べてトランスフォーマベースモデルの性能が高いことを少なくとも部分的に説明する可能性が高い。頻度ランクベースの予測方法は、最も低い予測（658個の正しいV遺伝子；3.5%）を有し、これは、データベースから軽鎖をランダムに選ぶ（915個の正しいV遺伝子；4.9%）ことよりも悪かった。

20

【0104】

データベース探索法は最も正確であり、4940個の正しいV遺伝子を有した（26.3%、図6A）。しかしながら、図6Aに見られるように、データベース探索法は、単一細胞ブラインドテストでは、テストセットでのようには上手く機能しなかった。実際に、トランスフォーマベースのモデルは、全ての単一細胞ブラインドテストでデータベース探索法よりも性能が優れていた。テストセットにおけるデータベース探索法の高性能は、大方、ブラインドセットでは当てはまらない、トレーニングセット及びテストセット内のクローン性関連物（clonal relatives）の存在に起因すると考えられる。実際に、単一B細胞クローンのメンバは、トレーニングセット及びテストセットにわたり分割された（表5参照）。例えば、予測が行われた18803個の重-軽鎖対のうち、897個はトレーニングセットと同一の重鎖配列を有した。接合アミノ酸配列において1つまでの不一致を許容する場合、テストセット内で、トレーニングセット内の重鎖と関連する3178個の配列が見つかった。これらの中で、729個（23%）は同一の軽鎖配列とペアリングされた。これらは、データベース探索法により正確にペアリングされる可能性がより高く、その一方で、トレーニングセットにおける重鎖レパトリの冗長性は、NLPに触発されたモデル（NLP-inspired models）のタスクを複雑にする恐れがある（言語Aでの複数の文章が言語Bでの単一の文章に翻訳される言語Aから言語Bへの翻訳を翻訳モデルが学習する必要があることと似る）。テストデータを使用した評価におけるこのバイアスは、クローンをトレーニングセット及びテストセットにわたり分割しないことにより、即ち、同じB細胞クローン（クラスタ）のメンバがトレーニングセットとテストセットとに分けられないことを保証することにより軽減することができる。

30

40

【0105】

50

【表 5】

表5. トレーニングセット及びテストセット内のクローナル関連配列の例

データセット	重 V 遺伝子 及び J 遺伝子	重鎖接合部	軽 V 遺伝子 及び J 遺伝子	軽鎖接合部	ドナー
トレーニング	IGHV1- 18:IGHJ6	CARATGAL YYYMDVW	IGLV1- 51:IGLJ2	CGAWDSSL SVVVF	ドナー1, DeKosky ら,2015
テスト	IGHV1- 18:IGHJ6	CARATGAL YYYMDVW	IGLV1- 51:IGLJ2	CGTWDSSL SVVVF	ドナー1, DeKosky ら,2015
トレーニング	IGHV4- 34:IGHJ3	CARGRGQ GGYPGLFV W	IGKV3- 20:IGKJ2	CQQYGSSP SYTF	ドナー2, DeKosky ら,2015
テスト	IGHV4- 34:IGHJ3	CARGRGLG GYPGLFVW	IGKV3- 20:IGKJ2	CQQYGSSP SYTF	ドナー2, DeKosky ら,2015

10

【0106】

20

したがって、テストセットでの評価は、そのような探索を実行するのに利用可能な対になった配列データの量が、BCRレパトリーの予期される多様性を真に表す（現実で該当するにはほど遠い）場合のみ現実的であるデータベース探索法の性能の歪んだ見方を与える。換言すれば、ブラインドテストでの評価は、方法の性能比較のはるかに現実的な見方を与える。

【0107】

トレーニングデータセット及びテストデータセットにわたるクローン性関連物の影響を除外するために、本発明者らは、Paired OASデータベースからの単一細胞データセットからの既知の対になった配列を使用してMatchmakerを *in silico* で検証した [Eccles et al., 2020; King et al., 2021; Setliff et al., 2019]。これらのブラインドテストデータセットのいずれもトレーニングセットとの間に共有する個人が存在しないため、あらゆる重複重鎖配列は、同じB細胞クローンのメンバではなく公開配列である可能性が高い。Eccles et al. [2020]及びSetliff et al. [2019]からのデータセット内の重鎖配列はいずれもトレーニングセットと一致せず、一方、King et al. [2021]のデータセット内の16個の重鎖配列はトレーニングセットと同一であった。重鎖配列の0.05%の重複は、重鎖収束の従来の観測に広く一致した [Briney et al., 2019]。

30

【0108】

3つの単一細胞データセットにわたり、Matchmakerの性能はトップであり、最高で9.8%の重鎖配列が正しい軽鎖V遺伝子と共に予測された（図6A）。Matchmakerは、King、Eccles、及びSetliffのそれぞれのデータセットにおける105、7、及び23個の重鎖に対して正しい軽鎖V遺伝子、接合配列、及びJ遺伝子を予測することも可能であった（表4）。それと比較して、データベース探索法は59、4、及び6個の正しい予測を有した。GRU法はMatchmakerほど正確ではなかったが、それもなお、2つの単一細胞データセットでデータベース探索法よりも高性能であった。これらの結果は、深層学習モデルが、特に新規の重鎖配列が存在する状況において、軽鎖予測よりも有用であることができることを示している。結果は、他の手法よりも優れた深層学習モデルの利点が、V遺伝子配列のみを見る場合よりも全長配列を見る場合にはるかに衝撃的であることをさらに示している。

40

【0109】

50

したがって、図 6 A 並びに表 3 及び 4 のデータは、2つの機械学習ベースの方法（及びカスタム設計のデータベース探索法）が全て、従来技術（頻度ベースの探索）に似た方法よりもはるかに高性能であったことを示す。これらの中で、トランスフォーマベースの方法は、全てのブラインドテストデータセットにわたり最良の性能を示し、全長配列を見た場合、さらに衝撃的な性能利点を示した。両機械学習ベースの方法も、3つのブラインドテストデータセットのうちの2つでデータベース探索法よりも高性能であった。

【0110】

図 6 B は、King et al. [2021] のデータセットで予測された軽鎖接合配列のレーベンシュタイン距離分布を示す。これは、トランスフォーマベースの方法（Matchmaker）が、任意の他の方法よりも短い距離を有するより多くの予測を有することを示し、予測された軽鎖接合アミノ酸配列が全体的に正しくない場合であっても、任意の他の方法を用いるよりも近い傾向があることを示す。

10

【0111】

興味深いことに、King et al. [2021] のデータセットでは、Matchmaker は、トレーニングセットにはない 29 個の軽鎖配列予測を生成し、これらのうちの 2 つは正しい軽鎖 V 遺伝子を有した。しかしながら、これらの新規の軽鎖では、軽鎖接合配列の予測はあまりよくなかった。29 個の予測された軽鎖接合アミノ酸配列と真のアミノ酸配列との間の平均レーベンシュタイン距離は、8 アミノ酸であった。

【0112】

実験検証

20

実際の制限に起因して、上述した最良性能の機械学習ベースの方法（Matchmaker）及びデータベース探索法のみを実験で検証した。方法論の最終検証として、これらの方法を使用して、9 つの治療用抗体（表 6）で軽鎖を予測した。これら 9 つは、標的が十分に確立されており、重鎖が抗原結合に様々なレベルの影響を与えるため、選ばれた。これは、蛋白質構造データバンク（Protein Data Bank）で利用可能なネイティブ抗体の分子構造を調べることにより重鎖の埋没表面積を計算することにより定量化された。この情報は表中の 8 / 9 個の抗体でのみ利用可能であった。ヒトではなくマウス由来のセツキシマブでは、ANARCI [Dunbar and Deane., 2016] を使用して最も近いヒト V 遺伝子及び J 遺伝子を推測した。

【0113】

30

【表 6】

表6. 9つの既知の抗体についてのネイティブ軽鎖と予測された軽鎖との比較。

治療用 mAb	標的	ネイティブ	Matchmaker	データベース探索
トラスツズマブ	Her2	IGKV1-39 - IGKJ1	IGLV2-14 - IGLJ2	IGKV1-5 - IGKJ1
ペルツズマブ	Her2	IGKV1-33 - IGKJ1	IGKV1-39 - IGKJ1	IGKV1-8 - IGKJ3
セツキシマブ	EGFR	IGKV6-21 - IGKJ2	IGKV3-15 - IGKJ4	IGKV1-39 - IGKJ4
パニツムマブ	EGFR	IGKV1-33 - IGKJ4	IGKV1-33 - IGKJ2	IGKV1-5 - IGKJ1
ペンブロリズマブ	PD1	IGKV3-11 - IGKJ4	IGLV3-1 - IGLJ2	IGLV3-19 - IGLJ2
ニボルマブ	PD1	IGKV3-11 - IGKJ1	IGKV1-33 - IGKJ2	IGKV2-30 - IGKJ4
デュルバルマブ	PDL1	IGKV3-20 - IGKJ1	IGKV1-39 - IGKJ1	IGKV1-33 - IGKJ3
アテゾリズマブ	PDL1	IGKV1-12 - IGKJ1	IGKV3-20 - IGKJ3	IGKV3-20 - IGKJ2
ゲムツズマブ	CD33	IGKV1-5 - IGKJ1	IGKV3-20 - IGKJ2	IGLV2-14 - IGLJ1

40

50

【0114】

データベース探索からの9つ全ての軽鎖予測は、安定した抗体をもたらし、一方、9つのMatchmakerペアリングの1つは失敗し、1つは、最初、発現することができなかったため、検定されなかった（表7参照）。ニボルマブのMatchmakerでペアリングされた形態を除き、いずれの方法からの他の全ての抗体も、熱安定性の抗体を生成した。Matchmakerからの軽鎖（IGKV1-39）を有するペルツズマブは発現したが、ネイティブ軽鎖（IGKV1-33）と類似するにもかかわらず、ペアリングは標的に結合しなかった。しかしながら、予測とネイティブの軽鎖配列との間で異なる少数の位置が重鎖界面に存在する。例えば、ネイティブペルツズマブ軽鎖は、アラニンが位置L40に有し、一方、Matchmaker予測の軽鎖はアスパラギンを有した。したがって、両方法とも（データベース探索予測及びトランスフォーマベースの予測）、大半又は全ての事例で安定した対を生成することが可能であった。従来技術により触発された頻度ベースの方法（上述したように、ランダムデータベース探索よりも性能が悪かった）の比較的低い性能に基づいて、これは、頻度ベースの方法を使用して取得され得た結果よりも良好な結果である可能性が高い。

【0115】

【表7】

表7. ネイティブペアリングに対する熱安定性=同じ又は1℃差以内。

治療用 mAb	標的	ネイティブ	Matchmaker	データベース探索
トラスツズマブ	Her2	=	=	=
ペルツズマブ	Her2	=	テストせず	=
セツキシマブ	EGFR	=	=	=
パニツムマブ	EGFR	=	=	=
ペンブロリズマブ	PD1	=	=	=
ニボルマブ	PD1	=	低下	=
デュルバルマブ	PDL1	=	=	=
アテゾリズマブ	PDL1	=	=	=
ゲムツズマブ	CD33	=	=	=

【0116】

各抗原への結合について、間接的ELISAにより27個の抗体（ネイティブ9個、Matchmaker9個、データベース9個）の各々を検定した。予測のいずれも、真の配列と同一の配列を有さず、したがって、いずれもオリジナル抗体の完全な結合親和性を回復しなかった。これらの治療用抗体の配列は親和性最適化（時にはヒト由来の配列に基づかない）に向けての集中的な遺伝子操作の結果であり、一方、予測方法は自然に生じる配列に基づくため、これは予期されていた。しかしながら、Matchmakerペアリングの抗体は、3/9の事例で結合の兆候を示し、一方、データベース探索により作られた抗体は1/9で結合を示した（表8、図8）。したがって、データは、Matchmaker法が、今まではBCRレパートリーのバルクシーケンシングから機能的抗体を提供するプロセスにおける大きなボトルネックであった工程である、さらなる遺伝子操作に向けて良好なベースを形成するのに十分な結合親和性を示す対の大部分を生成可能なことを示す。

【0117】

さらに、ここで使用される治療用抗体等の変異及び／又は遺伝子操作された抗体配列がいかに関与するかを反映した追加のトレーニングデータを使用して、この方法の性能をさらに強化することができると予期される。換言すれば、トレーニングデータを拡大して、モデルがより多くのデータから及び／又は最適化された配列を含むデータから学習できるようにすることは、ネイティブ抗体及び遺伝子操作された抗体の機能的ペアリングを予測するモデルの能力をさらに改善すると予期される。

【0118】

【表8】

表8. ネイティブペアリングに対する結合。

治療用抗体	標的	ネイティブ	Matchmaker	データベース探索	相互作用へのVH 寄与
トラスツズマブ	Her2	+++++	-	-	+
ペルツズマブ	Her2	+++++	-	++++	+++++
セツキシマブ	EGFR	+++++	+	-	++
パニツムマブ	EGFR	+++++	-	-	++
ペンブロリズマブ	PD1	+++++	-	-	++
ニボルマブ	PD1	+++++	+	-	+++
デュルバルマブ	PDL1	+++++	-	-	++
アテゾリズマブ	PDL1	+++++	++	-	++++
ゲムツズマブ	CD33	+++++	-	-	構造が利用 可能ではなく、 特定できない

10

【0119】

20

興味深いことに、Matchmakerは、パニツムマブに対して正しい軽鎖V遺伝子を予測したが、合成された抗体は結合を示さなかった。これは、正しいV遺伝子を予測することであっても、必ずしもBCRの機能を救うわけではないことを強調する。実際に、結合を示したMatchmaker又はデータベース探索法により予測されたペアリングのいずれも、正しいV遺伝子を予測しなかった。機能的抗体対の決定要因はあまり理解されていない。上述したように、V遺伝子は結合領域の大部分を形成するため、正しいV遺伝子の同定は、ペアリングが安定している可能性が高いことの指示及び少なくとも、バインダを同定するための見込みのある開始点を提供する。しかしながら、正しくない接合部はペアリングの最終的な結合親和性に大きく影響し得、わずかに異なるV遺伝子が存在する場合であっても結合に繋がり又は逆に、正しいV遺伝子が存在する場合であっても結合しないことに繋がる。これらの結果は、図6A及び図6B並びに表4における結果と組み合わせて、本明細書に記載の機械学習ベースの手法が、軽鎖配列の異なる部分間の複雑な相互作用（厳密なV遺伝子配列一致が、より適切な接合配列と組み合わせられた非厳密な一致ほど良好な結合親和性に繋がらないことがある）に関与する安定性及び結合の根底にある複雑な特徴を学習可能なことを示す。この能力は、完全ではないが、実験で評価された場合、新たに記載された非深層学習ベースの方法（データベース探索）よりも明らかに高性能である。さらに、*in-silico*での調査からの結果は、本明細書に記載の機械学習ベースの手法が、現行技術水準の頻度ベースの手法よりもさらに大差で高性能であることを示す。

30

【0120】

40

Matchmakerからの予測の中で最強のバインダは、データベース軽鎖とペアリングされたペルツズマブ重鎖であった。しかしながら、ペルツズマブの相互作用は主に重鎖により駆動され、これは、任意の安定した軽鎖が結合を可能にし得ることを示し得る。逆に、ニボルマブ及びセツキシマブのMatchmakerの軽鎖は、重鎖の重要度が比較的低いにもかかわらず、バインダに繋がった。

【0121】

本明細書に記載の方法の臨床有用性のさらなる検証及び証拠として、上述したMatchmakerモデルを使用して、免疫反応に関与し、コロナウイルスのスパイクタンパク質に結合する可能性が高いものとしてCOVID-19患者で同定された重鎖配列の重-軽鎖ペアリングを同定した。特に、Galson et al. (2020)のデータから選択されたこれら

50

の性質を有するCOVID-19患者から同定された18個の重鎖配列が、入力として、軽鎖とペアリングする方法に提供された。ペアリングを発現させ、ホモジニアス時間分解蛍光測定法（HTRF）を使用して武漢株スパイク抗原への結合について検定した。18個全ての抗体は発現に成功し、18個のうちの9個の抗体はスパイクタンパク質に結合するものとして確認された（結合はデルタF%>100として定義され、デルタFは、ネガティブコントロールと比べた蛍光リードアウトの割合増大を表す）。したがって、Matchmaker予測の50%は機能的抗体に繋がる。

【0122】

タンデムトランスフォーマモデル

本明細書に記載のGRUベース及びトランスフォーマベースのモデルの構造は、対になった重-軽鎖データがトレーニングに必要であるため、利用可能なトレーニングデータ量に幾つかの制限を課す。このタイプのデータは比較的限られた量で利用可能であり、完全な配列の代わりにV及びJ遺伝子識別子を提供することにより、その内容がさらに制限される（それにより、これらの分節の生殖細胞系配列に対する予測を実質的に制限する）。これらの制限を回避するために、対になっていない重鎖配列及び軽鎖配列のはるかに大きなデータセット（それぞれ4230万及び1530万個の配列）で各々がトレーニングされる2つのトランスフォーマを含むモデルを構築した。この結果を表9に示す。V遺伝子予測結果はMatchmakerよりもわずかに低かったが、正しいV遺伝子、接合アミノ酸配列、及びJ遺伝子を有する正しい予測の数はより高かった（表9と表5、6とを比較）。このデータは、この機械学習ベースの方法がまた、従来技術に似た方法（頻度ベースの探索）よりも高性能であることを示す。データは、タンデムトランスフォーマ手法が、少なくとも、トランスフォーマベースの手法が他の方法と比べて衝撃的な性能利点を既に示した全長配列にわたる性能を見る場合、トランスフォーマベースの方法をさらに改善する潜在性を有することをさらに示す。この手法の性能は、追加のトレーニングデータ及びトレーニング時間（ここでは実用的な理由で制限された）を用いてさらに増大する可能性が高い。

【0123】

【表9】

表9. 正しい軽鎖V遺伝子予測数

	タンデム AntiBERTa の正しい V	タンデム AntiBERTa の正しい V 遺伝子+接合アミノ酸 配列(CDRL3) +J 遺伝子	合計
Eccles	52 (7.8%)	6	741
King	2559 (8.4%)	180	30327
Setliff	380 (7.7%)	37	4944

【0124】

考察

この例は、BCR重-軽鎖ペアリングの問題に対する機械学習によるNLPに触発された手法を説明する。特にBCR重-軽鎖ペアリングの問題をNMTタスクと見なす2つのアーキテクチャ：Matchmaker、Seq2Seqトランスフォーマモデル及びSeq2Seq GRUベースのモデルについて説明する。知る限りでは、本研究は、入力として重鎖配列のみを必要とする深層人工ニューラルネットワーク（ANN）、特にSeq2SeqモデルのBCR重鎖ペアリングへの最初の適用である。Matchmakerは、この目的でのトランスフォーマモデルの最初の適用である。本明細書に記載の深層学習ベースの手法は、バルク軽鎖シーケンシングの必要性もなくしながら、BCRレパートリーを可能な限り深くカバーするという恩益を提供する。さらに、本手法は、トレーニングデータのセットからペアリングの一般特徴を学習し、この学習を使用して、以前は見られなかった鎖

のペアリングを予測することが可能である。逆に、データベース探索手法等の手法は、既知のペアリングに存在しないか又は既知のペアリングに存在する鎖からより離れているクエリ鎖を見る場合、すぐに機能停止しがちである。これは、BCRレパートリーの極めて大きな多様性を考慮して、多くの場合で有利である可能性が高いが、特に個人での所望の表現型の根底にあり得る特定の抗体又は他の希な抗体の同定等の用途の状況で有利である可能性が高い。

【0125】

本手法は、幾つかの代替の戦略と突き合わせて基準に従って評価された。重－軽鎖対のヘルドアウトテストセットでは、両深層ANNとも（特にMatchmaker）、かなり高性能であったが、データベース探索法が最高の正確性であった。同じB細胞クローンの関連物がトレーニングセットとヘルドアウトテストセットにわたって見つかったことを考慮すると、データベース探索法が最高の正確性であったことは驚くことではない。実際に、重－軽鎖対の十分に大きな知識ベースの取得が、現実的な見通しであったならば、正しい軽鎖相手方の同定を助けるであろう。しかしながら、ブラインドテストは、関連する重鎖配列が利用可能ではない場合、データベース探索の正確性がよくないことを示した。頻度探索は、全てのテスト状況にわたり最悪の性能であり、ランダム選択でさえわずかに良好であることがわかった。これは予想よりも低く、ペアリングの手段として軽鎖頻度を使用することの限界を裏付ける。

【0126】

ブラインドテストでの興味深い観測は、Matchmakerが完全に新規の軽鎖を予測可能であったこと、即ち、予測された軽鎖配列がトレーニングデータになかったことであった。これは、真の配列へのレーベンシュタイン距離がかなり低い（距離＝4）1つの新規の軽鎖を予測したGRUモデルでも当てはまった。29／29785個の重鎖のみがMatchmakerによる新規の軽鎖予測を有したが、これは、深層学習モデルが、トレーニングセットを記憶するのは対照的に、アテンションメカニズムにより重－軽鎖ペアリングの幾つかのルールを学習したことを示唆する。完全な重鎖配列及び軽鎖配列で構成されたより高品質のデータを用いれば、Matchmaker等の深層ANNモデルが、重－軽鎖ペアリングの改善された特定を学習することが予期される。それらのルールを考慮すれば、Matchmakerは、この研究で観測されたものよりも高い正確性の他の新規の軽鎖を生成するはずである。

【0127】

実験で検証された9つのポジティブコントロール治療用抗体のセットで、Matchmakerは、データベース探索法よりも多くのバインダを生成した。いずれの戦略から予測された抗体も、ネイティブ抗体ほど強力ではなく、大半の予測はバインダに繋がらず、問題の難しさを際立たせた。結合強度のばらつきは、結合への重鎖寄与を反映するように見えたが、Matchmakerはこれらの差により影響されるようには見えなかった。軽鎖と結合との間の精密な関係は不明瞭なままであるが、Matchmakerが結合抗体を生成するという事実は、さらなる抗体遺伝子操作機会に向けた出発点を与える。換言すれば、本方法は、重鎖のある割合に対してバインダを予測することが可能であり、少なくとも以下を理由として、従来技術からの大きな改善を表す：（a）入力として重鎖配列のみを必要とし、（b）入力として使用することができる配列及びデータセットのタイプに関して制限されず（即ち、クローン優勢試料内の高度過剰配列のみならず、任意のタイプの配列に対して有用な予測を提供可能であると予期される）、（c）他の方法よりも高いバインダ予測ヒット率を有する（現行技術水準よりも高い予測性能を有することが示された本明細書に新たに記載された方法でさえも）。そうして取得に成功した任意のそのようなバインダは、従来技術からの改善を表し、さらなる親和性改善に向けて有望な開始点として使用することができる。COVID-19患者からの重鎖配列を使用したさらなる検証は、Matchmakerの性能が、テストされた治療用抗体の限られたセットで生成されたペアリングの50%（9／18）が標的への幾らかの結合を示したため、このデータに基づいて示されたものよりもさらに高い可能性があることを示す。比較する方法と比べて観測された改善

10

20

30

40

50

は、それ自体で重要であるのみならず、BCRレパートリーの多様性を考慮して、機能的な重鎖-軽鎖ペアリングの同定が、さらなる開発に向けて十分に実行可能な候補に辿り着かせるいかなる改善も実用において極めて重要であるような極めて複雑な問題であることを考慮した場合、特に重要である。これは、探索効率の小さな改善であってさえも（そしてこれは決して小さな改善ではない）より高速の結果及び／又は治療用抗体の発見、テスト、最適化、及び開発における無駄な労力及びリソースの低減に繋がる、被験者からの試料からの治療用抗体発見等の用途の状況で特に当てはまり得る。さらに、本モデルは、トランスフォーマベースのモデルのトレーニングに典型的に使用されるデータセット（通常、数百万のオダの配列）と比べてかなり小さな153889個の重鎖-軽鎖対のトレーニングセットを使用してトレーニングされた。

10

【0128】

Matchmakerの実施態様への幾つかの可能な改善が考えられる。例えば、上述したモデルは、貪欲にトークンを予測する（即ち、一度に1つのトークン、換言すれば、個々の各位置で最も可能性の高いトークンが予測される）。ビームサーチ [Sustkever et al., 2014] 等の複数（例えば、3つ）の位置を同時に考慮する戦略を使用することは、モデルがより多くの解を探求し、したがって、最適未満の解に「足止め」される可能性がより低いため、予測正確性を上げるのに役立つはずである。これは、上述したタンデムトランスフォーマモデルに使用される手法である。さらに、先に例示したMatchmaker及びGRUモデルは、重複した固定サイズのk-merと遺伝子識別子との組合せをBCR配列のエンコードに使用する。単一アミノ酸又はアミノ酸のバイト対をトークンとして使用することは、モデルをさらに改善すると予期される。例えば、上述したタンデムトランスフォーマモデルは、各アミノ酸に単一のトークンを使用した。バイト対エンコードを使用すれば、データ駆動的に自動的に定義されるディクショナリを使用して、幾つかの非重複アミノ酸がトークンとしてエンコードされる。

20

【0129】

トレーニングセットの制限は、全長の重鎖配列及び軽鎖配列がないことである。実際に、そのようなデータは現在、限られた量でしか利用できず、したがって、J遺伝子識別子、V遺伝子識別子、及び接合配列のみを含むより大きなデータセットが使用された。代替のトークン化方式は、このギャップに対処するように設計されたが、効果的にV及びJ遺伝子の生殖細胞系軽鎖配列のみを予測する（それがトレーニングデータで利用可能な情報の全てであるため）。CDRL1及びCDRL2ループはCDRH3との接触を形成することができる [Leem et al., 2016] ため、予測された軽鎖は、*in vivo*でCDRH3との極めて重要な相互作用を形成する任意の変異アミノ酸を見逃すことがある。これは、トレーニングに異なる対になった重鎖-軽鎖データを使用し、例えば、10x Genomicsプラットフォームから等の細胞バーコーディングに基づくデータ等の全長配列を提供することにより解消し得る。これは、対になっていない全長配列のより大きなトレーニングデータセット（チェックポイントモデルをトレーニングするため）と潜在的により狭い配列カバレッジを有するより小さな対になったデータセットとの組合せを使用して、モデルをトレーニングすることができる上述したタンデムトランスフォーマ手法を使用して対処することもできる。

30

40

【0130】

バルク重鎖レパートリー解析を抗体発見に適用するために、軽鎖ペアリングの問題は該当するままである。本明細書に記載の深層学習ベースの手法は、入力として重鎖のみに依存することにおいて独自である。特に、Matchmaker（トランスフォーマベース）モデルは、複数のメトリックに基づいて最高の*in silico*正確性を有し、機能的抗体の生成について*in vitro*で検証された。対になった軽鎖情報が利用可能ではない場合に軽鎖を予測するために、これによりMatchmakerは独自に軽鎖を予測することができる。したがって、この手法は、軽鎖ペアリング情報におけるギャップを埋め、治療用抗体発見及び免疫系のよりよい理解を可能にする潜在性を有する。

【0131】

50

最後に、本手法について重鎖クエリに対する軽鎖ペアリングを同定する状況で説明したが、軽鎖クエリに対する重鎖ペアリングを同定する逆の問題にも適用可能である。重鎖シーケンシングはより一般的であり、重鎖は、特異性及び親和性を決めるに当たりより重要な役割を果たすと考えられるため、これはあまり頻度の高い問題ではない。

【0132】

引用文献

- Vander Heiden et al., 2017. Dysregulation of B Cell Repertoire Formation in Myasthenia Gravis Patients Revealed through Deep Sequencing. *J Immunol.* 2017 Feb 15;198(4):1460-1473.
- Bashford-Rogers et al, 2019. Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature* volume 574, pages122-126(2019).
- Nielsen et al., 2020. Human B Cell Clonal Expansion and Convergent Antibody Responses to SARS-CoV-2. *bioRxiv*. Preprint. 2020 Jul 9. doi: 10.1101/2020.07.08.194456.
- Simonich et al., 2019. Kappa chain maturation helps drive rapid development of an infant HIV-1 broadly neutralizing antibody lineage. *Nature Communications* volume 10, Article number: 2190 (2019).
- Krawczyk et al., 2019. Looking for therapeutic antibodies in next-generation sequencing repositories. *mAbs*. Volume 11, 2019 - Issue 7, Pages 1197-1205.
- Galson et al., 2020. Deep Sequencing of B Cell Receptor Repertoires From COVID-19 Patients Reveals Strong Convergent Immune Signatures. *Front. Immunol.*, 15 December 2020. doi.org/10.3389/fimmu.2020.605170.
- Mora and Walczak, 2019. How many different clonotypes do immune repertoires contain? *Current Opinion in Systems Biology*. Volume 18, December 2019, Pages 104-110
- Kovaltsuk et al., 2018. Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. *J Immunol* October 15, 2018, 201 (8) 2502-2509.
- Tiller et al., 2013. A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. *MAbs*. 2013 May 1; 5(3): 445-470.
- Tepliyakov et al., 2016. Structural diversity in a human antibody germline library. *MAbs*. Aug-Sep 2016;8(6):1045-63.
- Glanville et al., 2009. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *PNAS* December 1, 2009 106 (48) 20216-20221.
- Jayaram et al., 2012. Germline VH/VL pairing in antibodies. *Protein Engineering, Design and Selection*, Volume 25, Issue 10, October 2012, Pages 523-530.
- Ling et al., 2018. Effect of VH-VL Families in Pertuzumab and Trastuzumab Recombinant Production, Her2 and FcγIIA Binding. *Front. Immunol.*, 12 March 2018. doi.org/10.3389/fimmu.2018.00469
- DeKosky et al., 2013. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature Biotechnology* volume 31, pages166-169(2013).
- DeKosky et al., 2015. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nature Medicine* volume 21, pages86-91(2015).
- DeKosky et al., 2016. Large-scale sequence and structural comparisons

- of human naive and antigen-experienced antibody repertoires. PNAS May 10, 2016 113 (19) E2636-E2645.
- King et al., 2021. Single-cell analysis of human B cell maturation predicts how antibody class switching shapes selection dynamics. Science Immunology 12 Feb 2021. Vol. 6, Issue 56, eabe6291
- Eccles et al., 2020. T-bet+ Memory B Cells Link to Local Cross-Reactive IgG upon Human Rhinovirus Infection. Cell Reports Volume 30, Issue 2, 14 January 2020, Pages 351-366.e7
- Setliff et al., 2019. High-Throughput Mapping of B Cell Receptor Sequences to Antigen Specificity. Cell Volume 179, Issue 7, 12 December 2019, Pages 1636-1646.e15 10
- Reddy et al., 2010. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. Nature Biotechnology volume 28, pages 965-969 (2010).
- Zhu et al., 2013. Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. PNAS. 2013 Apr 16;110(16):6470-5.
- Raybould et al., 2021. Public Baseline and shared response structures support the theory of antibody repertoire functional commonality. PLoS Comput Biol 17(3): e1008781. 20
- Rakocevic et al., 2021. The landscape of high-affinity human antibodies against intratumoral antigens. bioRxiv. 8 Feb 2021. doi.org/10.1101/2021.02.06.430058
- Vaswani et al., 2017. Attention Is All You Need. arXiv:1706.03762
- Devlin et al., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805
- Radford et al., 2019. Language Models are Unsupervised Multitask Learners. <https://openai.com/blog/better-language-models/>
- Liu et al., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 30
- Rothe et al., 2020. Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. arXiv:1907.12461
- Bahdanau et al., 2015. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473
- Cho et al., 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:1412.3555
- Child et al., 2019. Generating Long Sequences with Sparse Transformers. arXiv:1904.10509
- Xiong et al., 2020. On Layer Normalization in the Transformer Architecture. arXiv:2002.04745 40
- Dunbar and Deane, 2016. ANARCI: antigen receptor numbering and receptor classification. Bioinformatics. 2016 Jan 15;32(2):298-300.
- Sutskever et al., 2014. Sequence to Sequence Learning with Neural Networks. arXiv:1409.3215
- Rees, 2020. Understanding the human antibody repertoire. MAbs. Jan-Dec 2020;12(1):1729683.
- Leem et al., 2016. ABodyBuilder: Automated antibody structure prediction with data-driven accuracy estimation. MAbs. 2016 Oct;8(7):1259-1268.
- Ye et al., 2013. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W34-40. 50

Furcy, David. Koenig, Sven. "Limited Discrepancy Beam Search". IJCAI '05 : Proceedings of the 19th international joint conference on Artificial intelligence. July 2005 Pages 125-131.

Zhou, Rong. Hansen, Eric. "Beam-Stack Search: Integrating Backtracking with Beam Search". Conference: Proceedings of the Fifteenth International Conference on Automated Planning and Scheduling (ICAPS 2005), June 5-10 2005.

Wolf et al., 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771

Carter Jason A., Preall Jonathan B., Grigaityte Kristina, Goldfless Stephen J., Jeffery Eric, Briggs Adrian W., Vigneault Francois, Atwal Gurinder S. "Single T Cell Sequencing Demonstrates the Functional Role of $\alpha\beta$ TCR Pairing in Cell Lineage and Antigen Specificity". Frontiers in Immunology. Vol. 10. 2019, p. 1516. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. (2017) 8:14049.

Howie B, Sherwood AM, Berkebile AD, Berka J, Emerson RO, Williamson DW, et al. High-throughput pairing of T cell receptor α and β sequences. Sci Transl Med. (2015) 7:301ra131.

Eve Richardson, Jacob D. Galson, Paul Kellam, Dominic F. Kelly, Sarah E. Smith, Anne Palser, Simon Watson & Charlotte M. Deane (2021) A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-pertussis toxoid antibodies, mAbs, 13:1.

Yi-Chun Hsiao, Yonglei Shang, Danielle M. DiCara, Angie Yee, Joyce Lai, Si Hyun Kim, Diego Ellerman, Racquel Corpuz, Yongmei Chen, Sharmila Rajan, Hao Cai, Yan Wu, Dhaya Seshasayee & Isidro Hoetzel (2019) Immune repertoire mining for rapid affinity optimization of mouse monoclonal antibodies, mAbs, 11:4, 735-746.

Warszawski S, Borenstein Katz A, Lipsh R, Khmelnsky L, Ben Nissan G, Javitt G, et al. (2019) Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces. PLoS Comput Biol 15(8): e1007207.

Seeliger D, Schulz P, Litzenburger T, Spitz J, Hoerer S, Blech M, Enenkel B, Studts JM, Garidel P, Karow AR. Boosting antibody developability through rational sequence optimization. MAbs. 2015;7(3):505-15. doi: 10.1080/19420862.2015.1017695.

Mason, D.M., Friedensohn, S., Weber, C.R. et al. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. Nat Biomed Eng (2021).

Galson Jacob D., Schaetzle Sebastian, Bashford-Rogers Rachael J. M., Raybould Matthew I. J., Kovaltsuk Aleksandr, Kilpatrick Gavin J., Minter Ralph, Finch Donna K., Dias Jorge, James Louisa K., Thomas Gavin, Lee Wing-Yiu Jason, Betley Jason, Cavlan Olivia, Leech Alex, Deane Charlotte M., Seoane Joan, Caldas Carlos, Pennington Daniel J., Pfeffer Paul, Osbourn Jane. Deep Sequencing of B Cell Receptor Repertoires From COVID-19 Patients Reveals Strong Convergent Immune Signatures. Frontiers in Immunology. Vol. 11, 2020.

【0133】

本明細書に引用された全ての引用文献は全体的に、本明細書において、まるで個々の各

公開特許、又は特許、又は特許出願が特に個々に参照により全体的に援用されると示されているかのような程度まで、参照により本明細書に援用される。

【0134】

本明細書に記載の特定の実施形態は、限定ではなく例として提供されている。記載の組成物、方法、及び技術の使用の種々の変更及び変形が、記載の技術の範囲及び趣旨から逸脱せずに当業者に明らかになろう。本明細書における任意のサブタイトルは、便宜上のみ含まれ、本開示を限定するものとして決して解釈されるべきではない。

【0135】

本明細書に記載の任意の実施形態の方法は、コンピュータプログラムとして、又はコンピュータプログラム製品として、又はコンピュータで実行されると、上述した方法を実行するように構成されたコンピュータプログラムを有するコンピュータ可読媒体として提供し得る。

【0136】

文脈により別段のことが示される場合を除き、上述した特徴の説明及び定義は、本発明のいかなる特定の態様又は実施形態にも限定されず、記載された全ての態様及び実施形態に等しく適用される。

【0137】

本明細書及び特許請求の範囲を通して、以下の用語は、文脈により明らかに別段のことが示される場合を除き、本明細書で明確に関連付けられた意味をとる。本明細書で使用される「一実施形態では」という句は、同じ実施形態を指すことも可能であるが、必ずしも同じ実施形態を指すわけではない。さらに、本明細書で使用される「別の実施形態では」という句は、異なる実施形態を指すことも可能であるが、必ずしも異なる実施形態を指すわけではない。したがって、後述するように、本発明の種々の実施形態は、本発明の範囲又は趣旨から逸脱せずに容易に組み合わせることが可能である。

【0138】

本明細書及び添付の特許請求の範囲で使用される場合、単数形「1つの(a)」、「1つの(an)」、及び「その(the)」は、文脈が明らかに別段のことを示すときを除き、複数形を含むことに留意しなければならない。範囲は、本明細書では「約」ある特定の値から及び／又は「約」別の特定の値までとして表現され得る。そのような範囲が表現される場合、別の実施形態はそのある特定の値から及び／又はその他方の特定の値までを含む。同様に、値が先行する「約」の使用により概算として表現される場合、特定の値が別の実施形態を形成することが理解されよう。数値に関連した「約」という用語は任意選択的であり、例えば＋／－10%を意味する。

【0139】

以下の特許請求の範囲を含め、本明細書全体を通して、文脈により別段のことが必要になる場合を除き、「備える(comprise)」、「含む」という言葉並びに「備える(comprises)」、「備えている」、及び「含んでいる」等の変形は、述べられた完全体、若しくはステップ、又は完全体若しくはステップの群の包含を暗示するが、他の任意の完全体、若しくはステップ、又は完全体若しくはステップの群の除外を暗示しないと理解されよう。

【0140】

本発明の他の態様及び実施形態は、文脈により別段のことが示される場合を除き、「備えている」という用語を「からなる」又は「基本的に～からなる」という用語で置換した上述した態様及び実施形態を提供する。

【0141】

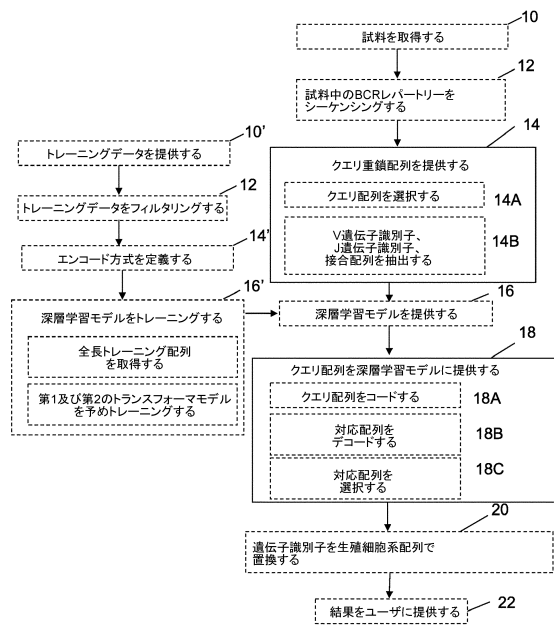
「及び／又は」は、本明細書で使用される場合、2つの指定された特徴又は構成要素の各々の、他方あり又は他方なしでの特定の開示として解釈されるべきである。例えば、「A及び／又はB」は、まるで各々が本明細書に個々に記載されるかのように、(i) A、(ii) B、及び(iii) A及びBの各々の特定の開示として解釈されるべきである。

【0142】

特定の形態で又は開示される機能を実行する手段又は開示される結果を得るための方法若しくはプロセスに関して適宜表現される、上記説明、又は以下の特許請求の範囲、又は添付図面に開示される特徴は、多様な形態で本発明を実現するために別個に又はそのような特徴の任意の組合せで利用することが可能である。

【図面】

【図 1】



【図 2】

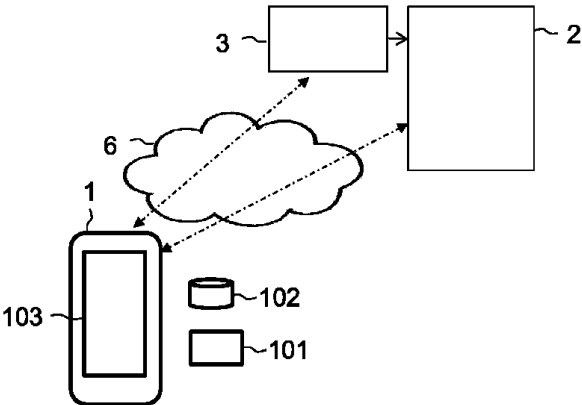
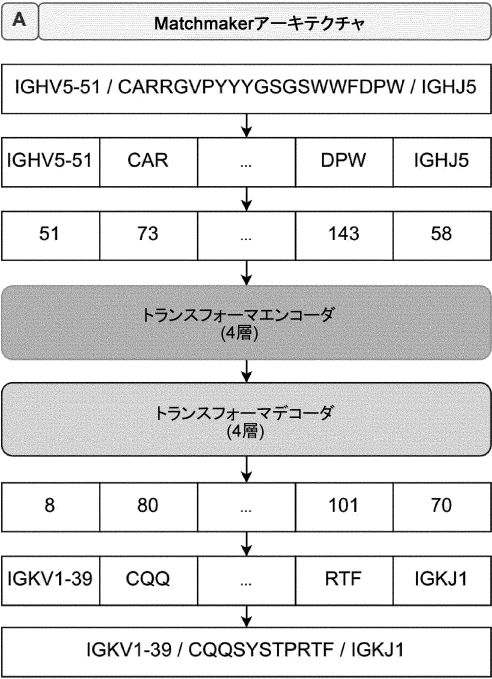
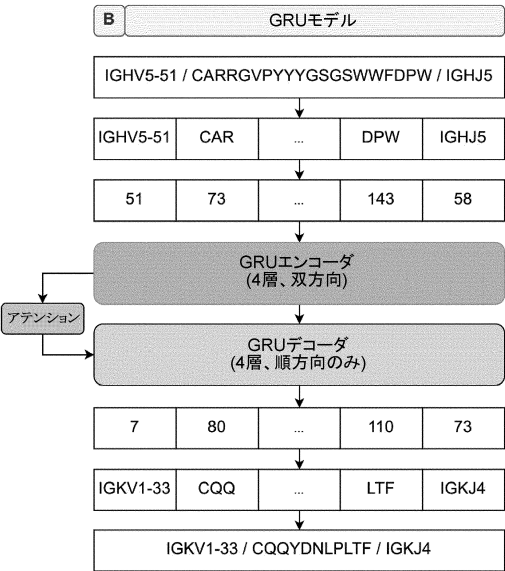


Fig. 2

【図 3 A】

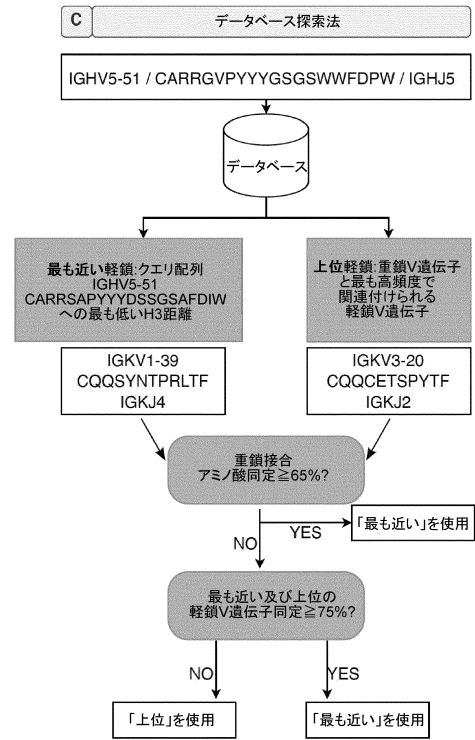


【図 3 B】



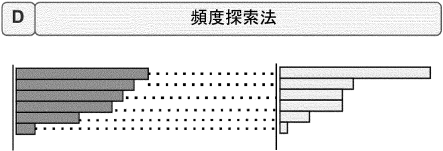
(続き)

【図 3 C】



(続き)

【図 3 D】



(続き)

10

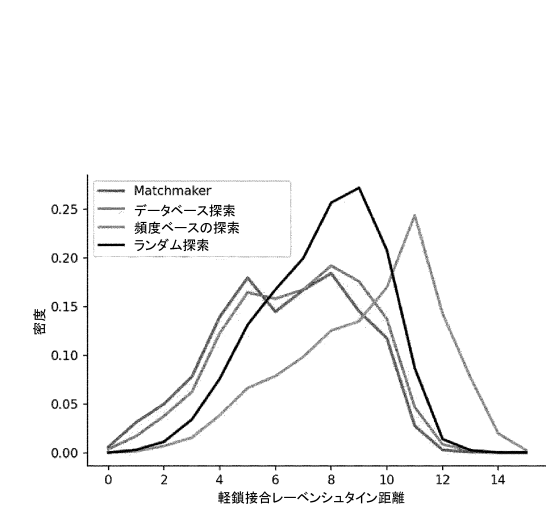
20

30

40

50

【図 6 B】

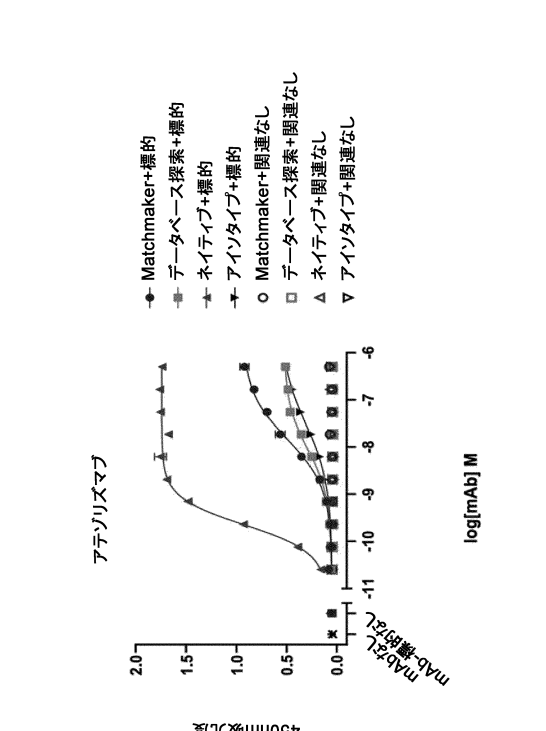


【図 7】

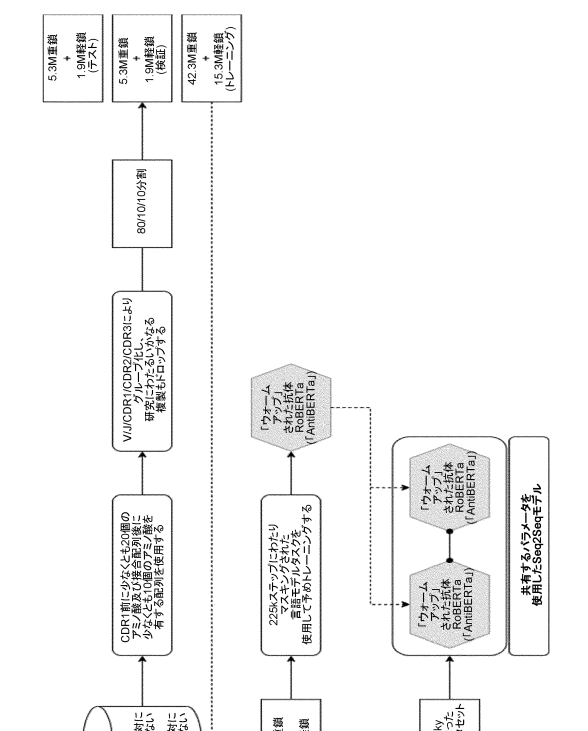


Fig. 7

【図 8】



【図 9】



【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2022/060073

A. CLASSIFICATION OF SUBJECT MATTER INV. G16B35/10 G16B40/20 ADD.		
According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) G16B		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) EPO-Internal, BIOSIS, COMPENDEX, EMBASE, INSPEC, WPI Data		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2020/236839 A2 (JUST BIOTHERAPEUTICS INC [US]; AMIMEUR TILELI [US] ET AL.) 26 November 2020 (2020-11-26) abstract paragraphs [0010] - [0012], [0023] - [0026], [0034], [0049], [0050], [0070], [0071], [0080], [0103], [0112], [0118] - [0121] paragraphs [0163] - [0165], [0283], [0284] ----- -/--	1-26
☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.		
* Special categories of cited documents : "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family		
Date of the actual completion of the international search 22 June 2022		Date of mailing of the international search report 30/06/2022
Name and mailing address of the ISA/ European Patent Office, P.B. 5618 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016		Authorized officer Thumb, Werner

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

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2022/060073

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	SAKA KOICHIRO ET AL: "Antibody design using LSTM based deep generative model from phage display library for affinity maturation", SCIENTIFIC REPORTS , vol. 11, no. 1 12 March 2021 (2021-03-12), XP055876990, DOI: 10.1038/s41598-021-85274-7 Retrieved from the Internet: URL:https://www.nature.com/articles/s41598-021-85274-7.pdf abstract page 9, paragraph 10 - page 11, paragraph 4 page 1, paragraph 1-4 -----	1-26
A	Ji Yanrong ET AL: "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome", bioRxiv, 19 September 2020 (2020-09-19), XP055877228, DOI: 10.1101/2020.09.17.301879 Retrieved from the Internet: URL:https://www.biorxiv.org/content/10.1101/2020.09.17.301879v1.full.pdf [retrieved on 2022-01-10] abstract page 2, paragraph 5 - page 4, paragraph 2 -----	1-26

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2022/060073

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2020236839 A2	26-11-2020	AU 2020278675 A1	23-12-2021
		AU 2022202958 A1	26-05-2022
		CA 3141476 A1	26-11-2020
		CN 114303201 A	08-04-2022
		EP 3956896 A2	23-02-2022
		KR 20220011148 A	27-01-2022
		WO 2020236839 A2	26-11-2020

フロントページの続き

MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,N
E,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,
CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,IT,JM,JO,J
P,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,N
A,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,ST,SV,SY,TH,TJ,
TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,WS,ZA,ZM,ZW

イギリス国, エヌ１シー ４エーエックス ロンドン キングス クロス イースト サイド オフィス
１．０２ アルケマブ セラピューティクス リミテッド

(72)発明者 ギャルソン, ジェイコブ

イギリス国, エヌ１シー ４エーエックス ロンドン キングス クロス イースト サイド オフィス
１．０２ アルケマブ セラピューティクス リミテッド

【要約の続き】

【選択図】 図１