(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)公表番号
特表2024-505638
(P2024-505638A)
(43)公表日　令和6年2月7日(2024.2.7)

(51)国際特許分類		FI			テーマコード（参考）
Ｇ１６Ｂ	*15/30 (2019.01)*	Ｇ１６Ｂ	15/30		４Ｂ０６３
Ｇ１６Ｂ	*40/20 (2019.01)*	Ｇ１６Ｂ	40/20		
Ｃ１２Ｑ	*1/6886(2018.01)*	Ｃ１２Ｑ	1/6886	Ｚ	

審査請求　未請求　予備審査請求　未請求　（全62頁）

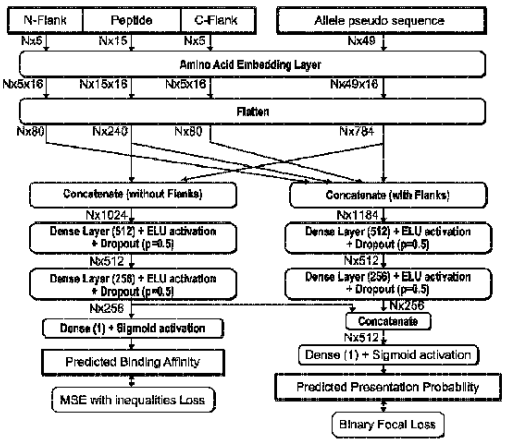
(21)出願番号　特願2023-543396(P2023-543396)
(86)(22)出願日　令和3年12月1日(2021.12.1)
(85)翻訳文提出日　令和5年9月1日(2023.9.1)
(86)国際出願番号　PCT/US2021/061399
(87)国際公開番号　WO2022/159176
(87)国際公開日　令和4年7月28日(2022.7.28)
(31)優先権主張番号　63/139,074
(32)優先日　令和3年1月19日(2021.1.19)
(33)優先権主張国・地域又は機関
米国(US)
(81)指定国・地域　AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,
最終頁に続く

(71)出願人　506329306
アマゾン テクノロジーズ インコーポレイテッド
アメリカ合衆国 ９８１０８－１２２６ ワシントン州 シアトル ピーオー ボックス ８１２２６
(74)代理人　100114890
弁理士　アインゼル・フェリックス＝ライン ハルト
(74)代理人　100098501
弁理士　森田 拓
(74)代理人　100116403
弁理士　前川 純一
(74)代理人　100134315
弁理士　永島 秀郎
最終頁に続く

(54)【発明の名称】　腫瘍特異的ネオアンチゲンのＭＨＣクラスＩまたはクラスＩＩ免疫原性を予測するためのディープラーニングモデル

(57)【要約】
本明細書では、腫瘍特異的ネオアンチゲンのＭＨＣクラスＩまたはＭＨＣクラスＩＩ免疫原性の予測方法であって、ＭＨＣクラスＩまたはＭＨＣクラスＩＩ結合親和性を予測することと、腫瘍特異的ネオアンチゲンがＭＨＣクラスＩまたはクラスＩＩタンパク質によって細胞表面上に提示されることになる可能性を予測することとを共に行うことによる、上記予測方法が開示される。

FIG. 1



【特許請求の範囲】
【請求項１】
腫瘍特異的ネオアンチゲンのＭＨＣクラスＩ免疫原性の予測方法であって、
ａ）腫瘍特異的ネオアンチゲンのペプチド配列及び前記ペプチド配列の対応する隣接領域を得ること；前記ペプチド配列及び前記隣接領域を数ベクトルにコード化することであって、各数ベクトルは、前記腫瘍特異的ネオアンチゲンの前記ペプチドをコードするアミノ酸残基及び前記隣接領域のアミノ酸残基、ならびに前記アミノ酸残基の位置を含む、前記コード化することと、
ｂ）ＨＬＡ対立遺伝子擬似配列を得ることであって、前記ＨＬＡ対立遺伝子擬似配列はＨＬＡ対立遺伝子を表す、前記得ることと；前記ＨＬＡ対立遺伝子配列を対応する数ベクトルにコード化することと、
ｃ）ニューラルネットワークモデルを使用して、前記腫瘍特異的ネオアンチゲンのＭＨＣクラスＩ結合親和性と、各腫瘍特異的ネオアンチゲンに関する、対応するペプチドがＭＨＣクラスＩタンパク質によって細胞表面上に提示されることになる数値で表した確率とを共に予測することであって、前記ニューラルネットワークモデルは、
（ｉ）前記ニューラルネットワークモデルをトレーニングデータセットに対してトレーニングして、前記ニューラルネットワークモデルのパフォーマンスを最適化することであって、前記トレーニングデータセットは、ペプチド－ＭＨＣクラスＩ親和性測定データセット及び細胞表面ペプチド提示データセットを含む、前記最適化することと、
（ｉｉ）前記腫瘍特異的ネオアンチゲンの前記ペプチド配列及び前記隣接領域を含む前記数ベクトル、ならびに前記ＨＬＡ対立遺伝子擬似配列層を含む前記数ベクトルを含む入力層と、
（ｉｉｉ）前記腫瘍特異的ネオアンチゲンの前記ペプチド配列及び前記隣接領域を含む前記数ベクトルならびに前記ＨＬＡ対立遺伝子擬似配列を含む前記数ベクトルをアミノ酸埋め込み層にコード化することと、
（ｉｖ）前記アミノ酸埋め込み層を平坦化して、前記腫瘍特異的ネオアンチゲンの各ペプチド配列、及び前記ペプチド配列の隣接領域、ならびに前記ＨＬＡ対立遺伝子擬似配列の数ベクトル表現を作成することと、
（ｖ）前記腫瘍特異的ネオアンチゲンのペプチド配列と前記ＨＬＡ対立遺伝子擬似配列を連結し、１つ以上の層及び／または１つ以上の活性化関数を適用することによって、前記腫瘍特異的ネオアンチゲンのＭＨＣクラスＩ結合親和性を予測することであって、出力は、前記腫瘍特異的ネオアンチゲンのＭＨＣクラスＩ結合親和性を表す数値スコアである、前記予測することと、
（ｖｉ）対象の前記ペプチド配列、前記ペプチド配列の隣接領域、及び前記ＨＬＡ対立遺伝子擬似配列を単一の数ベクトル中で連結し、１つ以上の層及び／または１つ以上の活性化関数を適用することによって、前記腫瘍特異的ネオアンチゲンがＭＨＣクラスＩタンパク質によって細胞表面上に提示されることになる確率を予測することであって、出力は、ペプチドがＭＨＣクラスＩタンパク質によって細胞表面上に提示されることになる数値で表した確率である、前記予測することと、を含む、前記予測することと、を含み、前記腫瘍特異的ネオアンチゲンのＭＨＣクラスＩ結合親和性、及び前記腫瘍特異的ネオアンチゲンがＭＨＣクラスＩタンパク質によって細胞表面上に提示されることになる前記数値で表した確率は、腫瘍特異的ネオアンチゲンのＭＨＣクラスＩ免疫原性の代用値である、前記予測方法。
【請求項２】
（ｉ）１つ以上のランキング指標を免疫原性検証データセットに適用することと、
（ｉｉ）ペプチドの予測されるＭＨＣクラスＩ結合親和性と、ペプチドがＭＨＣクラスＩタンパク質によって細胞表面上に提示されることになる前記数値で表した確率に基づいて、前記免疫原性検証データセット中の各対立遺伝子に関してペプチドをランキングすることと、
（ｉｉｉ）すべての対立遺伝子に関して前記１つ以上のランキング指標を集計すること

と
によって、前記ニューラルネットワークモデルを検証することをさらに含む、請求項１に記載の方法。
【請求項３】
前記１つ以上のランキング指標が、重み付けされた対立遺伝子頻度を使用することによって集計される、請求項２に記載の方法。
【請求項４】
前記ニューラルネットワークモデルが、汎対立遺伝子モデル、対立遺伝子特異的モデル（ａｌｌｅｌｅ－ｓｐｅｃｉｆｉｃ　ｍｏｄｅｌ）、スーパータイプ特異的モデル（ｓｕｐｅｒ－ｔｙｐｅ　ｓｐｅｃｉｆｉｃ　ｍｏｄｅｌ）、またはそれらの組み合わせである、先行請求項のいずれか１項に記載の方法。
【請求項５】
ＨＬＡ擬似配列の長さが約３０アミノ酸～約６０アミノ酸である、先行請求項のいずれか１項に記載の方法。
【請求項６】
前記腫瘍特異的ネオアンチゲンのペプチド配列が約８アミノ酸の長さ～約１５アミノ酸の長さである、先行請求項のいずれか１項に記載の方法。
【請求項７】
前記隣接領域が前記腫瘍特異的ネオアンチゲンのペプチド配列のすぐ左側、及び／または前記腫瘍特異的ネオアンチゲンのペプチド配列のすぐ右側である、先行請求項のいずれか１項に記載の方法。
【請求項８】
前記隣接領域の長さが約１０アミノ酸である、先行請求項のいずれか１項に記載の方法。
【請求項９】
前記腫瘍特異的ネオアンチゲンのすぐ左側である前記隣接領域の長さが約５アミノ酸である、請求項８に記載の方法。
【請求項１０】
前記腫瘍特異的ネオアンチゲンのすぐ右側である前記隣接領域の長さが約５アミノ酸である、請求項８に記載の方法。
【請求項１１】
前記ニューラルネットワークモデルのキャリブレーションを行うことをさらに含む、請求項１に記載の方法。
【請求項１２】
前記ニューラルネットワークモデルが、ポジティブトレーニングデータ及びネガティブトレーニングデータに対してトレーニングされる、請求項１に記載の方法。
【請求項１３】
ネガティブトレーニングデータが、腫瘍特異的ネオアンチゲンのＭＨＣクラスＩ結合親和性を有さない、及び／またはＭＨＣクラスＩタンパク質によって細胞表面上に提示されないペプチドを含む、請求項１２に記載の方法。
【請求項１４】
前記ＨＬＡ対立遺伝子が、ＨＬＡ－Ａ型、Ｂ型、またはＣ型である、先行請求項のいずれか１項に記載の方法。
【請求項１５】
前記腫瘍特異的ネオアンチゲンのＭＨＣクラスＩ免疫原性がＣＤ８＋Ｔ細胞免疫原性である、先行請求項のいずれか１項に記載の方法。
【請求項１６】
ＭＨＣクラスＩ免疫原性であると予測される１つ以上の腫瘍特異的ネオアンチゲンが免疫原性組成物用に選択される、先行請求項のいずれか１項に記載の方法。
【請求項１７】

少なくとも約２０個の腫瘍特異的ネオアンチゲンが前記免疫原性組成物用に選択される、請求項１６に記載の方法。
【請求項１８】
前記１つ以上の層が全結合層である、先行請求項のいずれか１項に記載の方法。
【請求項１９】
前記１つ以上の層がドロップアウト層である、先行請求項のいずれか１項に記載の方法。
【請求項２０】
前記１つ以上の層及び／または活性化関数が、１つ以上の全結合層を適用すること、ドロップアウト層を適用すること、及び活性化関数を適用することを含む、先行請求項のいずれか１項に記載の方法。
【請求項２１】
前記ニューラルネットワークモデルは、確率計算によって校正される、請求項１１に記載の方法。
【請求項２２】
確率計算が対象の対立遺伝子の全体的な提示の確率を推定する、請求項１１に記載の方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本出願は、２０２１年１月１９日出願の米国仮出願第６３／１３９，０７４号の利益を主張し、上記出願のすべての内容は本明細書に援用される。
【背景技術】
【０００２】
がんは世界的に主要な死因であり、全死亡者の４人に１人を占めている。Ｓｉｅｇｅｌ ｅｔ ａｌ．，ＣＡ： Ａ Ｃａｎｃｅｒ Ｊｏｕｒｎａｌ ｆｏｒ Ｃｌｉｎｉｃｉａｎｓ，６８：７－３０（２０１８）。２０１８年には１，８１０万人が新たにがんに罹患し、９６０万人ががん関連で死亡した。Ｂｒａｙ ｅｔ ａｌ．，ＣＡ： Ａ Ｃａｎｃｅｒ Ｊｏｕｒｎａｌ ｆｏｒ Ｃｌｉｎｉｃｉａｎｓ，６８（６）：３９４－４２４。切除技法（例えば、外科的処置及び放射線照射）ならびに化学的技法（例えば、化学療法剤）を含む、既存のがん治療における標準治療が多数存在する。残念ながら、かかる治療法は、多くの場合、重大なリスク、有毒な副作用、及び極めて高額な費用、ならびに有効性に対する不確実性を伴う。
【０００３】
がん免疫療法（例えば、がんワクチン）は、有望ながん治療法として登場している。がん免疫療法の目標は、免疫系を利用して、正常組織を無傷の状態に維持しつつ、がんを選択的に破壊することである。従来のがんワクチンは通常、腫瘍関連抗原を標的とする。腫瘍関連抗原は通常、正常組織中に存在するが、がんにおいて過剰発現する。しかしながら、これらの抗原は正常組織中に存在することが多いことから、免疫寛容により免疫の活性化が妨げられる場合がある。腫瘍関連抗原を標的としたいくつかの臨床試験では、標準治療と比較して、持続的で有益な効果を実証することができないでいる。Ｌｉ ｅｔ ａｌ．，Ａｎｎ Ｏｎｃｏｌ．，２８（Ｓｕｐｐｌ １２）： ｘｉｉ１１－ｘｉｉ１７（２０１７）。
【０００４】
ネオアンチゲンはがん免疫療法にとって魅力的な標的となる。ネオアンチゲンは個々の特異性を有する非自己タンパク質である。ネオアンチゲンは腫瘍細胞ゲノムにおけるランダムな体細胞変異に由来し、正常な細胞の表面には発現しない。同書。ネオアンチゲンは腫瘍細胞上でのみ発現され、中枢性免疫寛容を誘導しないため、がんネオアンチゲンを標的とするがんワクチンには、中枢性免疫寛容の低下及び安全性プロファイルの向上などの潜在的な利点を有する。同書。

【０００５】
がんの突然変異の状況は複雑であり、腫瘍の突然変異は概してそれぞれの個々の対象に固有である。配列決定によって検出される体細胞変異のほとんどは、有効なネオアンチゲンに繋がらない。有効であることが期待されるワクチンを設計するのに十分な精度で転写、翻訳され、腫瘍特異的ネオアンチゲンへとプロセッシングされるのは、腫瘍ＤＮＡまたは腫瘍細胞における突然変異のわずかな割合のみである。さらに、すべてのネオアンチゲンが免疫原性であるわけではない。実際、内因性ネオアンチゲンを自発的に認識するＴ細胞の割合は約１％～２％である。Ｐａｕｌ　ｅｔ　ａｌ．，　Ｊ．　Ｉｍｍｕｎｏｌ．，１９２，　５８３１－５８３９（２０１３）；Ｙｅｗｄｅｌｌ，　Ｉｍｍｕｎｉｔｙ，２５，　５３３－５４３（２００６）を参照のこと。上記約１％のＭＨＣに結合するネオアンチゲンのうち、Ｔ細胞によって認識されることになるのは約５０％に過ぎず、自然にプロセッシングされて、腫瘍細胞を死滅させることが可能になるのは３０～４０％に過ぎない。同書。

【０００６】
現在のインシリコ法は主として、いずれのネオアンチゲンペプチドがＭＨＣ－ＩもしくはＭＨＣ－ＩＩ分子に結合するかをモデル化すること、またはいずれのネオアンチゲンが腫瘍細胞によって短鎖ペプチドへとプロセッシングされ、ＭＨＣクラスＩ／ＩＩ分子によって提示される可能性が高いかを予測することにのみ焦点を当てている。利用可能なツールは、上記提示されたペプチドのいずれが免疫原性であるかを判定する予測精度に欠ける。そのため、既存の方法では陽性的中率が低くなる。例えば、ある研究において、３人の黒色腫患者がそれぞれ、インビトロで確認されたＭＨＣ結合親和性が５００ｎＭ未満である７種のペプチドで免疫化された。Ｃａｒｒｅｎｏ　ｅｔ　ａｌ．，　Ｓｃｉｅｎｃｅ，３４８，　８０３－８０８（２０１５）。試験した２１種のペプチドのうち、Ｔ細胞応答を誘導したのは９種のみであった。同書。ネオアンチゲンペプチドを含有する個別化ワクチンが陽性的中率の低い方法を使用して設計されている場合、患者が投与される治療用ネオアンチゲンが、当該のがんに対する免疫応答を誘発する能力を有する可能性は低い。

【０００７】
したがって、がん患者における個別化ネオアンチゲンを系統的に識別することは、個別化がんワクチンの開発を成功させるための重要な要件である。よって、個別化ワクチン用に陽性的中率が高い免疫原性ネオアンチゲン候補を効率的に且つ高精度で予測することは依然として課題である。

【０００８】
発明の概要
本開示は、腫瘍特異的ネオアンチゲンのＭＨＣクラスＩまたはＭＨＣクラスＩＩ免疫原性の新規な予測方法であって、ＭＨＣクラスＩまたはＭＨＣクラスＩＩ結合親和性を予測することと、上記腫瘍特異的ネオアンチゲンがＭＨＣクラスＩまたはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示されることになる可能性を予測することとを共に行うことによる、上記方法に関する。本方法は、腫瘍細胞によって、対象のＭＨＣクラスＩまたはＭＨＣクラスＩＩ分子に結合し、且つＴ細胞受容体と接触し、最終的には免疫原性となる可能性が高いペプチドへとプロセッシングされる可能性が高い、高的中率の腫瘍特異的ネオアンチゲンを高精度で識別する。この方法は、免疫応答を誘発することになるネオアンチゲンを識別するための高い予測精度を有し、このことは有効な個別化免疫原性組成物（例えば、がんワクチン）を開発するために重要である。これは既存の手法において障害となっていた。

【０００９】
さらに、本明細書に記載の方法は、ゴールドスタンダード予測子であるＭＨＣｆｌｕｒｒｙ－１．４結合親和性予測子、ならびにＭＨＣｆｌｕｒｒｙ－２．０予測子よりもパフォーマンスが優れている。さらなる詳細については「実施例」の節を参照されたい。これらのモデルのそれぞれは別個の予測子である。ＭＨＣｆｌｕｒｒｙ－１．４は対立遺伝子特異的なＭＨＣクラスＩ結合予測因子である（Ｏ’Ｄｏｎｎｅｌｌ　ｅｔ　ａｌ．，　Ｃｅ

ｌｌ　Ｓｙｓｔｅｍ，　７：１２９－１３２（２０１８））。ＭＨＣｆｌｕｒｒｙ－２．０予測因子はＭＨＣクラスＩを介して提示されるペプチドの汎対立遺伝子（ｐａｎ－ａｌｌｅｌｅ）予測因子である。

【００１０】

本発明者らはさらに、方法の検索能力を直接評価する独自のベンチマークを作成するための免疫原性評価データセットを開発しており、上記データセットは、本方法の、所与のＭＨＣクラスＩ対立遺伝子に対する大規模なペプチド候補のプールから免疫原性ペプチドを検索する能力を直接評価する。この能力は個別化免疫療法に基づくワクチン設計にとって必須の要素である。

【００１１】

上記腫瘍特異的ネオアンチゲンのＭＨＣクラスＩ免疫原性の予測方法をより明確に説明するために、図１に本方法の概略的な流れ図を示す。

【００１２】

上記腫瘍特異的ネオアンチゲンのＭＨＣクラスＩまたはＭＨＣクラスＩＩ免疫原性の予測方法は、腫瘍特異的ネオアンチゲンのペプチド配列及び上記ペプチド配列の対応する隣接領域を得ることから開始される。上記隣接領域は、上記腫瘍特異的ネオアンチゲンペプチドのすぐ左側、または上記腫瘍特異的ネオアンチゲンペプチドのすぐ右側のアミノ酸配列であってよい。例えば、上記隣接領域は、上記腫瘍特異的ネオアンチゲンペプチドのＣ末端及び／またはＮ末端上にあるアミノ酸配列であってよい。通常、上記隣接領域の長さは約１０アミノ酸であってよい。例えば、上記腫瘍特異的ネオアンチゲンのすぐ左側の隣接領域の長さは約５アミノ酸であってよい。例えば、上記腫瘍特異的ネオアンチゲンのすぐ右側の隣接領域の長さは約５アミノ酸であってよい。次いで、上記ペプチド配列及び上記隣接領域が数ベクトルにコード化される。各数ベクトルは、上記腫瘍特異的ネオアンチゲンの上記ペプチドをコードするアミノ酸残基及び上記隣接領域、ならびに上記アミノ酸残基の位置を含む。ＨＬＡ対立遺伝子を表すＨＬＡ対立遺伝子擬似配列が得られる。上記ＨＬＡ擬似配列の長さは少なくとも約２０～約１００アミノ酸であってよい。好ましくは、上記ＨＬＡ擬似配列の長さは少なくとも約３０～６０アミノ酸の長さである。上記ＨＬＡ対立遺伝子擬似配列は対応する数ベクトルにコード化される。上記ＨＬＡ対立遺伝子は、Ａ型、Ｂ型、またはＣ型、ＤＱ、ＤＰ、またはＤＲである。

【００１３】

次いで、ニューラルネットワークモデルが使用されて、腫瘍特異的ネオアンチゲンのＭＨＣクラスＩまたはＭＨＣクラスＩＩ結合親和性と、対象の各ペプチドに関する、対応するペプチドがＭＨＣクラスＩまたはＭＨＣによって細胞表面上に提示されることになる数値で表した確率とを共に予測する。上記ニューラルネットワークモデルは、汎対立遺伝子モデル（ｐａｎ－ａｌｌｅｌｅ　ｍｏｄｅｌ）、対立遺伝子特異的モデル（ａｌｌｅｌｅ－ｓｐｅｃｉｆｉｃ　ｍｏｄｅｌ）、スーパータイプ特異的モデル（ｓｕｐｅｒ－ｔｙｐｅ　ｓｐｅｃｉｆｉｃ　ｍｏｄｅｌ）、またはそれらの組み合わせであってよい。

【００１４】

初めに、上記ニューラルネットワークモデルを、トレーニングデータセットに対してトレーニングして、上記ニューラルネットワークモデルのパフォーマンスを最適化する。上記トレーニングデータセットは、ペプチド－ＭＨＣクラスＩまたはＭＨＣクラスＩＩ親和性測定データセット及び細胞表面ペプチド提示データセットを含む。上記ニューラルネットワークモデルは、ポジティブトレーニングデータ及びネガティブトレーニングデータに対してトレーニングされることが好ましい。上記ネガティブトレーニングデータは、腫瘍特異的ネオアンチゲンのＭＨＣクラスＩもしくはクラスＩＩ結合親和性を有さない、及び／またはＭＨＣクラスＩもしくはクラスＩＩタンパク質によって細胞表面上に提示されないペプチドを含んでいてもよい。

【００１５】

モデル入力層は、上記腫瘍特異的ネオアンチゲンの上記ペプチド配列及び上記隣接領域を含む数ベクトル、ならびにＨＬＡ対立遺伝子擬似配列を含む数ベクトルを含む。次に、

上記数ベクトルのそれぞれがアミノ酸埋め込み層にコード化される。次いで、上記ニューラルネットワークモデルは上記アミノ酸埋め込み層を平坦化し、上記腫瘍特異的ネオアンチゲンの各ペプチド配列及び上記ペプチド配列の隣接領域、ならびに上記ＨＬＡ対立遺伝子擬似配列の数ベクトル表現を作成する。

【００１６】

上記ペプチド腫瘍特異的ネオアンチゲンのＭＨＣクラスＩまたはＭＨＣクラスＩＩ結合親和性を予測するために、上記腫瘍特異的ネオアンチゲンのペプチド配列と上記ＨＬＡ対立遺伝子擬似配列が連結される。上記モデルは、１つ以上の層及び／または１つ以上の活性化関数を適用することをさらに含む。例えば、上記モデルは１つ以上の結合層を適用することを含んでいてもよい。例えば、上記モデルはドロップアウト層を適用することを含んでいてもよい。例えば、上記モデルは活性化関数を適用することを含んでいてもよい。場合によっては、上記モデルは１つ以上の結合層を適用すること、１つ以上のドロップアウト層を適用すること、及び／または活性化関数を適用することを含んでいてもよい。出力はペプチドリガンド－ＭＨＣクラスＩまたはＭＨＣクラスＩＩ結合親和性を表す数値スコアである。上記腫瘍特異的ネオアンチゲンがＭＨＣクラスＩまたはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示されることになる確率を予測するために、対象のペプチド配列、上記ペプチド配列の隣接領域、及び上記ＨＬＡ対立遺伝子擬似配列が単一の数ベクトル中で連結される。予測されるペプチドリガンド－ＭＨＣクラスＩまたはＭＨＣクラスＩＩ結合親和性も連結される。上記モデルは１つ以上の層及び／または１つ以上の活性化関数を適用することをさらに含む。例えば、上記モデルは１つ以上の結合層を適用することを含んでいてもよい。例えば、上記モデルはドロップアウト層を適用することを含んでいてもよい。例えば、上記モデルは活性化関数を適用することを含んでいてもよい。上記モデルは１つ以上の結合層を適用すること、ドロップアウト層を適用すること、及び／または活性化関数を適用することを含んでいてもよい。出力は、ペプチドがＭＨＣクラスＩまたはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示されることになる数値で表した確率である。上記腫瘍特異的ネオアンチゲンのＭＨＣクラスＩまたはＭＨＣクラスＩＩ結合親和性、及び上記腫瘍特異的ネオアンチゲンがＭＨＣクラスＩまたはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示されることになる数値で表した確率は、腫瘍特異的ネオアンチゲンのＭＨＣクラスＩ免疫原性の代用値である。一般に、ＭＨＣクラスＩ免疫原性はＣＤ８＋　Ｔ細胞免疫原性である。ＭＨＣクラスＩＩ免疫原性はＣＤ４＋Ｔ細胞免疫原性である。

【００１７】

本方法は、１つ以上のランキング指標を免疫原性検証データセットに適用することと、ペプチドの予測されるＭＨＣクラスＩ結合親和性と、ペプチドがＭＨＣクラスＩタンパク質によって細胞表面上に提示されることになる数値で表した確率に基づいて、上記免疫原性検証データセット中の各対立遺伝子に関してペプチドをランキングすることと、すべての対立遺伝子に関して上記ランキング指標を集計することとによって、上記ニューラルネットワークを検証することをさらに含んでいてもよい。上記ランキング指標は重み付けされた対立遺伝子頻度を使用することによって集計されてもよい。本方法は上記ニューラルネットワークのキャリブレーションを行うことをさらに含んでいてもよい。上記ニューラルネットワークモデルのキャリブレーションは確率計算によって行われてもよい。この計算により、対象の対立遺伝子の全体的な提示の確率を推定することができる。

【００１８】

ＭＨＣクラスＩまたはＭＨＣクラスＩＩ免疫原性であると予測される腫瘍特異的ネオアンチゲンが、免疫原性組成物用に選択されてもよい。通常、約１０～約２０種の腫瘍特異的ネオアンチゲンが免疫原性組成物用に選択されてもよい。

【図面の簡単な説明】

【００１９】

【図１】モデルアーキテクチャ図である。モデル入力：ａ）ペプチド配列（及び隣接領域）；ｂ）対立遺伝子擬似配列。モデル出力：ａ）予測される結合親和性；ｂ）予測される

提示の確率。モデルのトレーニングに使用される損失関数：ａ）結合親和性の予測に対しては不等号を伴うＭＳＥ損失（ＭＳＥ－ｗｉｔｈ－ｉｎｅｑｕａｌｉｔｉｅｓ　ｌｏｓｓ）；ｂ）提示の確率の予測に対しては二値（ｂｉｎａｒｙ）ｆｏｃａｌ　ｌｏｓｓ。

【図２】Ａは閾値を超える類似性を有するパートナーペプチドを持たないペプチドの割合を示すグラフである。Ｂは３つのデータセット（親和性、提示、及び免疫原性）にわたるペプチド長の分布を示すグラフである。

【図３】ペプチド－ＭＨＣ結合親和性ラベル分布を示すグラフである。Ｙ軸は各ビンに属するトレーニングデータ中のサンプルの割合を表す。

【図４】データセットサンプルにおけるＨＬＡ対立遺伝子スーパータイプの分布を示すグラフである。

【図５Ａ】ＨＬＡ対立遺伝子毎のサンプル分布を示すグラフであり、ペプチドＭＨＣ結合親和性のサンプル分布を示す。

【図５Ｂ】ＨＬＡ対立遺伝子毎のサンプル分布を示すグラフであり、細胞表面へのペプチド提示のサンプル分布を示す。

【図５Ｃ】ＨＬＡ対立遺伝子毎のサンプル分布を示すグラフであり、Ｔ細胞免疫原性のサンプル分布を示す。

【図６】米国集団における対立遺伝子頻度に基づく、免疫原性対立遺伝子頻度重みを示すグラフである。

【図７】異なる値のアルファ（各損失要素の重みを決定する）に伴うパフォーマンスの比較を示すグラフである。

【図８】Ａは免疫原性を有する割合と予測される提示の確率との間の相関性を示すグラフである。Ｂは免疫原性を有する割合と予測される結合親和性との間の相関性を示すグラフである。この実験は、グランドトゥルースの免疫原性ラベルが既知であり、汎モデルを使用して提示及び結合親和性が予測された免疫原性検証セットに対して実施された。

【図９】プロバイダネットワーク環境の例を示す図である。

【図１０】いくつかの実施形態に係る、ストレージサービス及びハードウェア仮想化を顧客に提供する例示的なプロバイダネットワークのブロック図である。

【図１１】例示的なコンピュータシステムを示すブロック図である。

【図１２】モデル入力を示す図である。モデルは２つのシーケンス、トークンシーケンスと対応するセグメントシーケンスを受け取る。上記トークンシーケンスは、＜ＣＬＳ＞トークン、対立遺伝子疑似配列トークン、＜ＳＥ［＞トークン、ｎ－隣接領域トークン、ペプチド、ｃ－隣接領域トークン、及び＜ＥＯＳ＞トークンを連結することによって構成される。上記セグメントシーケンスは、対応するトークンが属するセグメントを示すインデックスを提供する。

【図１３】マルチヘッドセルフアテンションモジュール及びそれに続くフィードフォワードモジュール（間に挟まれたＧＥＬＵ活性化関数を有する２つの線形層で構成される）で構成されるＴｒａｎｓｆｏｒｍｅｒ層の概略図である。層正規化は各モジュールの最初に適用され、残差ドロップアウトは各コンポーネントの最後で残差結合の前に適用される。

【発明を実施するための形態】

【００２０】

本開示は、腫瘍特異的ネオアンチゲンのＭＨＣクラスＩまたはＭＨＣクラスＩＩ免疫原性の新規な予測方法であって、ＭＨＣクラスＩまたはＭＨＣクラスＩＩ結合親和性のいずれかを予測すること、及び上記腫瘍特異的ネオアンチゲンがＭＨＣクラスＩまたはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示されることになる可能性を予測することを共に行うことによる、上記方法に関する。上記の新規な方法は、好ましくは、腫瘍特異的ネオアンチゲンのＭＨＣクラスＩまたはＭＨＣクラスＩＩ免疫原性を予測するために使用される。

【００２１】

本方法は、腫瘍特異的ネオアンチゲンの配列決定データ及びＨＬＡ対立遺伝子を表すＨＬＡ対立遺伝子擬似配列を得ることを含む。例えば、エクソーム、トランスクリプトーム

、及び／または全ゲノムのヌクレオチド配列決定を使用して、腫瘍特異的ネオアンチゲンの配列決定データ及びペプチド配列を得ることができる。本方法は、各腫瘍特異的ネオアンチゲンのペプチド配列及び任意選択で隣接領域を、対応する数ベクトルにコード化することをさらに含んでいてもよい。各数ベクトルには、ペプチド配列を構成するアミノ酸残基及び上記アミノ酸残基の位置を記述する情報が含まれる。

【００２２】

本方法は上記ＨＬＡ擬似配列を数ベクトルにコード化することも含んでいてよい。本方法は、ニューラルネットワークモデル中に上記数ベクトルを入力して、上記腫瘍特異的ネオアンチゲンのＭＨＣクラスＩまたはＭＨＣクラスＩＩ結合親和性と、各腫瘍特異的ネオアンチゲンの、対応するペプチドがＭＨＣクラスＩまたはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示されることになる数値で表した確率とを共に予測することを含んでいてもよい。これらの予測はどちらも、ＭＨＣクラスＩまたはＭＨＣクラスＩＩ免疫原性（例えば、ＣＤ８＋　Ｔ細胞免疫原性またはＣＤ４＋　Ｔ細胞免疫原性）を予測するための代用値として使用することができる。上記数ベクトルが上記ニューラルネットワークモデルに入力された後に、上記数ベクトルがアミノ酸埋め込み層に変換されてもよく、次いで上記アミノ酸埋め込み層が平坦化されて、それぞれの上記腫瘍特異的ネオアンチゲンのペプチド配列、任意選択で上記ペプチド隣接領域、及び上記ＨＬＡ対立遺伝子擬似配列の数ベクトル表現を作成してもよい。

【００２３】

次に、上記ニューラルネットワークモデルを使用して、上記腫瘍特異的ネオアンチゲンのＭＨＣクラスＩまたはＭＨＣクラスＩＩ結合親和性と、上記腫瘍特異的ネオアンチゲンがＭＨＣクラスＩまたはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示されることになる数値で表した確率とを予測することができる。これらの予測は、上記腫瘍特異的ネオアンチゲンのペプチド配列、上記ＨＬＡ対立遺伝子擬似配列、及び任意選択で上記ペプチド隣接領域を連結することによって実施することができる。次いで、１つ以上の層及び／または関数が適用されてもよい。例えば、１つ以上の全結合ｄｅｎｓｅ　ｌａｙｅｒが適用されてもよい。例えば、１つ以上のドロップアウト層が適用されてもよい。例えば、１つ以上の活性化関数が適用されてもよい。実施形態において、１つ以上の全結合ｄｅｎｓｅ　ｌａｙｅｒ、１つ以上のドロップアウト層、及び／または、１つ以上の活性化関数の組み合わせが適用されてもよい。出力は、上記腫瘍特異的ネオアンチゲンのＭＨＣクラスＩまたはＭＨＣクラスＩＩ結合親和性、及び／またはペプチドがＭＨＣクラスＩまたはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示されることになる数値で表した確率を表す数値スコアである。これらの予測値は免疫原性の代用値とすることができる。次いで、免疫原性腫瘍特異的ネオアンチゲンを、個別化された免疫原性組成物の含有物用に選択することができる。

【００２４】

本明細書で開示される予測はトレーニングデータセットに基づいて特定される。上記トレーニングデータセットは複数のサンプルを含む。上記トレーニングデータセットは、ペプチドのＭＨＣクラスＩまたはＭＨＣクラスＩＩ親和性測定データセット及び細胞表面へのペプチド提示データセットを含んでいてもよい。上記ニューラルネットワークモデルは、ポジティブトレーニングデータのみならず、ネガティブトレーニングデータに対してもトレーニングされることが好ましい。上記ネガティブトレーニングデータは、腫瘍特異的ネオアンチゲンのＭＨＣクラスＩもしくはＭＨＣクラスＩＩ結合親和性を有さない、及び／またはＭＨＣクラスＩもしくはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示されないペプチドを含んでいてもよい。

【００２５】

ＭＨＣクラスＩまたはＭＨＣクラスＩＩ免疫原性であると予測されるネオアンチゲンは、免疫原性組成物の含有物用に選択することができる。通常、約１０～約２０種の腫瘍特異的ネオアンチゲンを免疫原性組成物用に選択することができる。例えば、上記免疫原性組成物は、約５種、約６種、約７種、約８種、約９種、約１０種、約１１種、約１２種、

約１３種、約１４種、約１５種、約１６種、約１７種、約１８種、約１９種、約２０種、約２１種、約２２種、約２３種、約２４種、または約２５種の腫瘍特異的ネオアンチゲンを含んでいてもよい。

【００２６】

Ｉ．定義

この開示において引用されるすべての刊行物及び特許は、それらの全体が援用される。援用される資料が本明細書と矛盾するかまたは不整合である範囲内で、本明細書はいずれのかかる資料にも優先することになる。本明細書におけるいかなる参考文献の引用も、かかる参考文献が本開示に対する先行技術であることを認めるものではない。本明細書の態様に関連するさまざまな用語が、本明細書及び特許請求の範囲全体にわたって使用されている。かかる用語には、別段の指示がない限り、当該技術分野における通常の意味が与えられるべきものである。他の特に定義される用語は、本明細書において示される定義と整合する形態で解釈されるべきものである。

【００２７】

本明細書では、単数形「ａ」、「ａｎ」、及び「ｔｈｅ」は、文脈上明らかに別段の指示がない限り複数形を含む。「含む」、「など」などの用語は、特に別段の指示がない限り、制限のない包含を意味するものとする。

【００２８】

本明細書では、用語「がん」とは、細胞の集団が無制御の増殖、不死性、転移能、急速な増殖（ｇｒｏｗｔｈ）及び増殖（ｐｒｏｌｉｆｅｒａｔｉｏｎ）速度、ならびに／または特定の形態学的特徴によってキャラクタライズされる、対象における生理学的状態をいう。がんは腫瘍または塊の形態をとる場合が多いが、対象の体内に単独で存在する場合、または白血病細胞もしくはリンパ腫細胞などの独立した細胞として血流中を循環する場合もある。用語がんには、血液悪性腫瘍、固形腫瘍、肉腫、癌腫、及び他の固形腫瘍及び非固形腫瘍を含む、あらゆる種類のがんならびに転移が含まれる。がんの例としては、癌腫、リンパ腫、芽腫、肉腫、及び白血病が挙げられるが、これらに限定はされない。かかるがんのより詳細な例としては、扁平上皮癌、小細胞肺癌、非小細胞肺癌、肺の腺癌、肺の扁平上皮癌、腹膜の癌、肝細胞癌、胃腸癌、膵臓癌、神経膠芽腫、子宮頸癌、卵巣癌、肝臓癌、膀胱癌、肝細胞腫、乳癌（例えば、トリプルネガティブ乳癌、ホルモン受容体陽性乳癌）、骨肉腫、黒色腫、結腸癌、結腸直腸癌、子宮内膜（例えば、漿液性）または子宮癌、唾液腺癌、腎臓癌、肝臓癌、前立腺癌、外陰癌、甲状腺癌、肝癌、及び多様な種類の頭頸部癌が挙げられる。トリプルネガティブ乳癌とは、エストロゲン受容体（ＥＲ）、プロゲステロン受容体（ＰＲ）、及びＨｅｒ２／ｎｅｕの遺伝子の発現が陰性である乳癌を指す。ホルモン受容体陽性乳癌とは、ＥＲ及びＰＲのうち少なくとも１つが陽性であり、且つＨｅｒ２／ｎｅｕ（ＨＥＲ２）が陰性である乳癌をいう。

【００２９】

本明細書では、用語「ネオアンチゲン」とは、例えば、腫瘍細胞における突然変異または腫瘍細胞に特異的な翻訳後修飾を介して、対応する親抗原と区別される少なくとも１つの変化を有する抗原を指す。突然変異としては、フレームシフト、インデル、ミスセンスもしくはナンセンス置換、スプライス部位の変化、ゲノムの再配列もしくは遺伝子融合、またはネオアンチゲンを生じさせるいずれかのゲノム発現の変化を挙げることができる。突然変異としてはスプライス突然変異を挙げることができる。腫瘍細胞に特異的な翻訳後修飾としては異常なリン酸化を挙げることができる。腫瘍細胞に特異的な翻訳後修飾としてはプロテアソームによって生成するスプライス抗原を挙げることもできる。Ｌｉｐｅｔ ａｌ．，Ｓｃｉｅｎｃｅ，３５４（６３１０）：３５４：３５８（２０１６）を参照されたい。一般に、点変異は腫瘍における変異の約９５％を占め、インデル及びフレームシフト変異が残余を占める。Ｓｎｙｄｅｒ ｅｔ ａｌ．，Ｎ Ｅｎｇｌ Ｊ Ｍｅｄ．，３７１：２１８９－２１９９（２０１４）を参照されたい。

【００３０】

本明細書では、用語「腫瘍特異的ネオアンチゲン」とは、対象の腫瘍細胞または組織に

は存在するが、該対象の正常細胞または組織には存在しないネオアンチゲンである。
【００３１】
本明細書では、用語「免疫原性」とは、免疫応答（例えば、Ｔ細胞応答、Ｂ細胞応答、または両方）を誘発する能力をいう。
【００３２】
本明細書では、用語「ＨＬＡ対立遺伝子擬似配列」とは、ＨＬＡ対立遺伝子アミノ酸配列を表すためのアルゴリズムによって生み出されるアミノ酸配列を指す。
【００３３】
本明細書では、用語「対象」とは、ヒト、非ヒト霊長動物、げっ歯動物などを含む、但しこれらに限定されない任意の哺乳動物などの任意の動物をいう。いくつかの実施形態において、上記哺乳動物はマウスである。いくつかの実施形態において、上記哺乳動物はヒトである。
【００３４】
本明細書では、用語「腫瘍細胞」とは、がん細胞であるか、またはがん細胞に由来する任意の細胞を指す。用語「腫瘍細胞」は、がん様の特性、例えば、制御不能な増殖、抗増殖シグナルに対する抵抗性、転移能、及びプログラム細胞死を起こす能力の喪失を示す細胞を指す場合もある。
【００３５】
本明細書では、用語「ニューラルネットワーク」とは、通常は確率的勾配降下法及び逆伝播によってトレーニングされた、線形変換及びそれに続く要素毎の非線形性の多層からなる、分類または回帰のための機械学習モデルをいう。
【００３６】
本明細書において直接定義されていないいずれの用語も、本発明の技術分野内で理解される、当該用語に一般的に結びつく意味を有するものと解されるべきものである。本明細書において直接定義されていないいずれの用語も、本発明の技術分野内で理解される、当該用語に一般的に結びつく複数の意味を有するものと解されるべきものである。本明細書では、本発明の態様の組成物、デバイス、方法など、及びそれらの製造方法または使用方法の説明において、医師にさらなる指針を提供するために、特定の用語について論じる。同一の事物が複数の言い方で表される場合があることは理解されよう。したがって、本明細書で論じられる任意の１つ以上の用語に対して、それらに代わる言葉及び同義語が使用される場合がある。ある用語が本明細書において詳述されているまたは論じられているかどうかに重きを置くべきではない。いくつかの同義語または代替可能な方法、材料などが示される。１つもしくはいくつかの同義語または等価物が記載される場合、明示的に記載されていない限り、他の同義語または等価物の使用が排除されるものではない。用語の例を含む例を使用するのは、説明することのみを目的としており、本明細書における本発明の態様の範囲及び意味を限定するものではない。
【００３７】
本方法のさらなる説明及び本方法の実施のための指針が本明細書に提供される。説明を容易にするために、ＭＨＣクラスＩ免疫原性の予測であって、ＭＨＣクラスＩ結合親和性を予測すること、及び上記腫瘍特異的ネオアンチゲンがＭＨＣクラスＩタンパク質によって細胞表面上に提示されることになる可能性を予測することを共に行うことによる、上記予測の好ましい態様に関して、さらなる詳細及び指針が提供される。さらなる詳細及び指針は、ＭＨＣクラスＩＩ免疫原性の予測にも関連することが意図される。
【００３８】
ＩＩ．トレーニング
上記ニューラルネットワークモデルは、トレーニングデータセットに対して上記ニューラルネットワークモデルをトレーニングして、該ニューラルネットワークが、腫瘍特異的ネオアンチゲンのＭＨＣクラスＩまたはＭＨＣクラスＩＩ結合親和性と、上記腫瘍特異的ネオアンチゲンがＭＨＣクラスＩまたはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示されることになる確率とを予測することが可能になるように、上記ニューラルネッ

トワークモデルのパフォーマンスを最適化することを含んでいてもよい。

【００３９】

本明細書に記載の方法で使用されるトレーニングデータセットは、複数のサンプルを含む。上記トレーニングデータは様々なデータを含んでいてもよい。上記トレーニングデータは、ペプチドのＭＨＣクラスＩまたはＭＨＣクラスＩＩ親和性測定データセット及び細胞表面へのペプチド提示データセット、及び任意選択で免疫原性データセットを含むことができる。上記トレーニングデータはヒトライノウイルスデータを含んでいてもよい。ネガティブサンプルは免疫原性評価に使用することができる。上記トレーニングデータセットを使用して、１つ以上のニューラルネットワークモデルをトレーニングすることができる。実施形態において、１つ以上のニューラルネットワークモデルをトレーニングすることができる。例えば、少なくとも２つ以上のニューラルネットワークをトレーニングすることができる。少なくとも約３、４、５、６、７、８、９、１０以上のニューラルネットワークをトレーニングすることができる。

【００４０】

ペプチドのＭＨＣクラスＩまたはＭＨＣクラスＩＩ親和性測定データを含むデータセットは、特定のＭＨＣクラスＩ対立遺伝子またはＭＨＣクラスＩＩ対立遺伝子に対する実験的に測定された結合親和性ペプチドを含んでいてもよい。上記データセットは公開されているデータソースなどの１つ以上のデータソースから取得することができる。例えば、上記データセットはＩｍｍｕｎｅ　Ｅｐｉｔｏｐｅ　Ｄａｔａｂａｓｅ（“ＩＥＤＢ，”ｉｅｄｂ．ｏｒｇ）から取得することができる。上記トレーニングデータセットは１つ以上のデータソースに基づいてさらに拡張することができる。上記トレーニングデータセットは、本明細書に開示の方法のためにさらに精選されてもよい。例えば、上記トレーニングデータは、ペプチドと関連する各ＭＨＣ分子の間の結合親和性の予測を含んでいてもよい。上記データセット中の実験的に測定された結合親和性は、定量的（「＝」の不等式を伴う）、定性的（「＜」もしくは「＞」の不等式を伴う）、またはそれらの組み合わせであってもよい。定量的データはＩＣ５０（ｍＭ）値を含んでいてもよい。定性的データセットは、陽性－高（例えば、結合親和性＜１００ｎｍ）、陽性－中（例えば、結合親和性＜１，０００ｎｍ）、陽性－低（例えば、結合親和性＜５，０００ｎｍ）、または陰性（例えば、結合親和性＞５，０００ｎｍ）のように表されてもよい。ＭＨＣクラスＩまたはＭＨＣクラスＩＩ親和性測定データセットを含むトレーニングデータセットは、ＭＨＣから溶離し、質量分析法によって同定されたペプチドを含んでいてもよい。

【００４１】

ＭＨＣクラスＩ親和性測定データセットをさらに精選して、特定のＭＨＣクラスＩペプチド対立遺伝子に対する予測される結合親和性のサブセットを保持してもよい。例えば、ＨＬＡ－Ａ型、ＨＬＡ－Ｂ型、及び／またはＨＬＡ－Ｃ型対立遺伝子のエントリが保持されてもよい。例えば、特定の長さのペプチドが保持されてもよい。長さが少なくとも約５アミノ酸～約２０アミノ酸のペプチドが保持されてもよい。上記ペプチドの長さは、約５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、または２０アミノ酸であってよい。上記トレーニングセット中のペプチドの長さは同一であるかまたは異なっていてもよく、上記長さはＭＨＣ対立遺伝子の種類に応じて変化する場合がある。上記ペプチドの長さは約５～約１５アミノ酸であることが好ましい。翻訳後修飾または非標準アミノ酸を含むペプチドは除去されてもよい。

【００４２】

ＭＨＣクラスＩＩ親和性測定データセットをさらに精選して、特定のＭＨＣクラスＩＩペプチド対立遺伝子に対する予測される結合親和性のサブセットを保持してもよい。例えば、ＨＬＡ－ＤＰ、ＨＬＡ－ＤＱ、及び／またはＨＬＡ－ＤＲ対立遺伝子のエントリが保持されてもよい。例えば、特定の長さのペプチドが保持されてもよい。長さが少なくとも約５アミノ酸～約４０アミノ酸のペプチドが保持されてもよい。上記ペプチドの長さは、約５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０、２１、２２、２３、２４、２５、２６、２７、２８、２９、３０、３１、３２、

３３、３４、３５、３６、３７、３８、３９、または４０アミノ酸であってよい。上記トレーニングセット中のペプチドの長さは同一であるかまたは異なっていてもよく、上記長さはＭＨＣ対立遺伝子の種類に応じて変化する場合がある。通常、上記ペプチドの長さは約１３～約３５アミノである。翻訳後修飾または非標準アミノ酸を含むペプチドは除去されてもよい。

【００４３】

上記ＭＨＣクラスＩまたはＭＨＣクラスＩＩ親和性測定データは回帰モデルとして提供されてもよい。特に、損失関数が使用されてもよい。例示的な損失関数としては、交差エントロピー損失関数、平均二乗誤差、Ｈｕｂｅｒ損失、Ｋｕｌｌｂａｃｋ－Ｌｅｉｂｌｅｒ、ＭＡＥ（Ｌ１）、ＭＡＥ（Ｌ３）、尤度関数、及びヒンジ損失が挙げられる。特に、平均二乗損失関数の変化形が使用されてもよい。平均二乗損失関数はL_{BA-MSE}によって表すことができ、但し、測定値は（＞）または（＜）を伴い、データセット中の定量的及び定性的なペプチド－ＭＨＣ結合親和性測定値の両方の処理に関して、上記測定値は、不等式に反する場合にのみ損失に寄与する。使用することができる式は、

【数１】

$$L_{BA\text{-}MSE} = \frac{1}{N}\sum_{i=1}^{N} l(\hat{Y}_i, Y_i)$$

$$l(\hat{Y}_i, Y_i) = \begin{cases} (Y_i - \hat{Y}_i)^2 & \text{、測定値に関する不等式が（=）の場合} \\ \max(\hat{Y}_i - Y_i, 0)^2 & \text{、測定値に関する不等式が（<）の場合} \\ \max(Y_i - \hat{Y}_i, 0)^2 & \text{、測定値に関する不等式が（>）の場合} \end{cases}$$

スキーム１

である。

【００４４】

【数２】

Y_i及び$\hat{Y}_i$

はｉ番目のサンプルに関するペプチド－ＭＨＣ結合親和性の目標値及び予測値を表す。上記親和性の目標はトレーニングの前に変換することができる。［０，１］の範囲内の目標値に対して、［０，５００００］の範囲のＩＣ５０（ｎＭ）値。次の関数：

【数３】

$$Y_i = 1 - \frac{\log IC50_i}{\log 50000}$$

スキーム２

を使用してＩＣ５０（ｎＭ）値を変換することができる。

【００４５】

上記ＭＨＣクラスＩまたはＭＨＣクラスＩＩ親和性測定データセットは、少なくとも約５，０００種、約１０，０００種、約１５，０００種、約２０，０００種、約２５，０００種、約３０，０００種、約３５，０００種、約４０，０００種、約４５，０００種、約５０，０００種、約６０，０００種、約７０，０００種、約８０，０００種、約９０，０００種、約１００，０００種、約１５０，０００種、約２００，０００種、約２５０，０

００種、約３００，０００種、約３５０，０００種、約４００，０００種、約４５０，０００種、約５００，０００種、約５５０，０００種、約６００，０００種、約６５０，０００種、約７００，０００種、約７５０，０００種、約８００，０００種、約８５０，０００種、約９００，０００種、約９５０，０００種、約１，０００，０００種、約１，２５０，０００種、約１，５００，０００種、約１，７５０，０００種、約２，０００，０００種、もしくはそれを超えるペプチドの、ＭＨＣクラスＩまたはＭＨＣクラスＩＩペプチド対立遺伝子に対する結合親和性の測定値を含んでいてもよい。一般に、上記ＭＨＣクラスＩまたはＭＨＣクラスＩＩ親和性測定データセットは、少なくとも約２０，０００種の固有のペプチドを含む。

【００４６】

上記細胞表面へのペプチド提示データセットは、ＨＬＡ分子を介して提示されることが知られているペプチドを含んでいてもよい。細胞表面ペプチドは、例えば、ペプチド溶離実験または質量分析データによって測定することができる。上記細胞表面へのペプチド提示データセットは、公的に入手可能なデータソースなどの１つ以上のデータソースから取得することができる。例えば、Ｉｍｍｕｎｅ　Ｅｐｉｔｏｐｅ　Ｄａｔａｂａｓｅ　（“ＩＥＤＢ，”　ｉｅｄｂ．ｏｒｇ）またはＳｙｓｔｅＭＨＣプロジェクトで生み出されたペプチドは、有用なデータソースとなり得る。上記細胞表面へのペプチド提示データセットは、さらに実験的に作成されてもよい。例えば、ペプチドは、ＨＬＡペプチドを発現する細胞株由来のペプチドを希釈し、質量分析法によって該ペプチドを分析することによって調製することができる。上記トレーニングデータセットは、１つ以上のデータソースに基づいてさらに拡張することができる。

【００４７】

これらのトレーニングデータセットは通常、本明細書で開示の方法のために精選される。ペプチド配列は一般に文字列として表され、該文字列において、各文字はアミノ酸を表す。ペプチド配列は、当該ペプチドのアミノ酸及び該アミノ酸の位置を記述する情報を含む数ベクトルに変換されてもよい。上記数ベクトルは二項分類であってよい。例えば、k^i個のアミノ酸を有するペプチド配列p^iは、２０種のアミノ酸の行ベクトル（２０－ｋ）によって表され、該行ベクトルにおいて、当該ペプチド配列の特定の位置のアミノ酸のアルファベットに対応する単一の要素は１の値をもつことになる。残余の要素は０の値をもつことになる。例として、アミノ酸アルファベットがＡ、Ｃ、Ｄ、Ｅ、Ｆ、Ｇ、Ｈ、Ｉ、Ｋ、Ｌ、Ｍ、Ｎ、Ｐ、Ｑ、Ｒ、Ｓ、Ｔ、Ｖ、Ｗ、及びＹの場合、３個のアミノ酸のペプチド配列ＡＦＰは、６０個の要素の行ベクトル及び

【数４】

$$p^i = 1000000000000000000000001000000000000000000000000000001000000$$

によって表すことができる。上記トレーニングデータセットがさまざまな長さのアミノ酸配列を含む場合には、上記数ベクトルは、当該ペプチドを同じ長さにコード化するためのパディング文字を含んでいてもよい。上記パディング文字は、当該ペプチド配列の左側または当該ペプチド配列の右側に適用することができる。当業者であれば、他の種類の分類システムを適用することができることを認識しよう。

【００４８】

損失関数を上記細胞表面へのペプチド提示データセットに使用してもよい。損失関数の例としては、交差エントロピー損失関数、平均二乗誤差、Ｈｕｂｅｒ損失、Ｋｕｌｌｂａｃｋ－Ｌｅｉｂｌｅｒ、ＭＡＥ（Ｌ１）、ＭＡＥ（Ｌ３）、尤度関数、及びヒンジ損失が挙げられる。特に、上記細胞表面へのペプチド提示データセットによって、交差エントロピー損失関数が使用されてもよい。特定の実施形態において、Ｆｏｃａｌ　Ｌｏｓｓ二値分類が使用されてもよい。Ｆｏｃａｌ　Ｌｏｓｓ二値分類を使用して、上記データセットにおける不均衡を低減することができる。Ｆｏｃａｌ　ＬｏｓｓはL_{P-FL}として表す

ことができ、Ｌｐ－ＦＬは標準的な二値交差エントロピー損失を重み付けして拡張したもので、分類が不十分なサンプルをより強調する。Ｆｏｃａｌ　Ｌｏｓｓの式は以下のとおりである。

【数５】

$$L_{\text{p-FL}} = -\frac{1}{N}\sum_{i=1}^{N}(1-p_{ti})^{\gamma}\log p_{ti}$$

但し、正しいクラスp_{ti}は以下のとおりである。

$$p_{ti} = \begin{cases} \hat{Y}_i & , Y_i = 1 \\ 1-\hat{Y}_i & , Y_1 = 0 \end{cases}$$

スキーム３

γは実数パラメータであり、１に設定される。二値の場合、

【数６】

$$Y_i \in \{0, 1\}$$

はグランドトゥルースラベルであり、

【数７】

$$\hat{Y}_i$$

はｉ番目のサンプルに関する予測される提示の確率である。

【００４９】

ランキングオブジェクティブ（ｒａｎｋｉｎｇ　ｏｂｊｅｃｔｉｖｅ）をトレーニング用の上記細胞表面へのペプチド提示データセットにさらに使用してもよい。例えば、Ｎ－ｗａｙ分類をランキング指向のトレーニングに使用してもよい。Ｎ－ｗａｙ分類により、データセット中のポジティブサンプルとネガティブサンプルを競合させることが可能になる。次いで、上記サンプルを、Ｎ個のサンプルの各セットがポジティブサンプルであるサンプルとして分類することができる（Ｎ＝ネガティブサンプルの数＋１）。Ｎ－ｗａｙ分類損失の場合、交差エントロピー損失関数またはＦｏｃａｌ　Ｌｏｓｓ関数を適用することができる。

【００５０】

上記細胞表面へのペプチド提示データセットは、少なくとも約５，０００種、約１０，０００種、約１５，０００種、約２０，０００種、約２５，０００種、約３０，０００種、約３５，０００種、約４０，０００種、約４５，０００種、約５０，０００種、約６０，０００種、約７０，０００種、約８０，０００種、約９０，０００種、約１００，０００種、約１５０，０００種、約２００，０００種、約２５０，０００種、約３００，０００種、約３５０，０００種、約４００，０００種、約４５０，０００種、約５００，０００種、約５５０，０００種、約６００，０００種、約６５０，０００種、約７００，０００種、約７５０，０００種、約８００，０００種、約８５０，０００種、約９００，０００種、約９５０，０００種、約１，０００，０００種、約１，２５０，０００種、約１，５００，０００種、約１，７５０，０００種、約２，０００，０００種、またはそれを超えるペプチドを含んでいてもよい。３５，０００のサンプルを超えるトレーニングデータセットが好ましい。

【００５１】
上記ニューラルネットワークモデルのトレーニングは、すべてのトレーニングデータに対して、または一部のトレーニングデータに対して行うことができる。例えば、上記ニューラルネットワークモデルのトレーニングは、約１００％のトレーニングデータに対して、約９５％、約９０％、約８５％、約８０％、約７５％、約７０％、約６５％、約６０％、約５５％、またはそれを下回るトレーニングデータセットに対して行うことができる。上記ニューラルネットワークモデルのトレーニングは、すべての上記ＭＨＣクラスＩまたはＭＨＣクラスＩＩ親和性測定値セット中のトレーニングデータ、及びすべての上記細胞表面へのペプチド提示トレーニングデータセット中のトレーニングデータに対して行うことができる。例えば、上記ニューラルネットワークモデルのトレーニングは、約１００％、約９５％、約９０％、約８５％、約８０％、約７５％、約７０％、約６５％、約６０％、約５５％、もしくはそれを下回る上記ＭＨＣクラスＩまたはＭＨＣクラスＩＩ親和性測定データセット、及び／あるいは約１００％、約９５％、約９０％、約８５％、約８０％、約７５％、約７０％、約６５％、約６０％、約５５％、もしくはそれを下回る上記細胞表面へのペプチド提示トレーニングデータセットに対して行うことができる。

【００５２】
一実施形態において、１つ以上のトレーニングデータセットのトレーニングデータにｃｒｏｓｓ－ｔｒａｉｎｅｄを行ってもよい。例えば、上記ＭＨＣクラスＩまたはＭＨＣクラスＩＩ親和性測定データセット及び上記細胞表面へのペプチド提示データセットにｃｒｏｓｓ－ｔｒａｉｎｅｄを行ってもよい。各データセットは通常、単一の既知の目標を含む。例えば、上記ＭＨＣクラスＩまたはＭＨＣクラスＩＩ親和性測定データはペプチドの親和性を含み、上記細胞表面へのペプチド提示データセットは、ＭＨＣクラスＩまたはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示され可能性があるペプチドを含む。上記トレーニングデータセットのｃｒｏｓｓ－ｔｒａｉｎを行うために、各トレーニングを行うセットの目標を推測してもよい。例えば、ペプチドを細胞表面上に提示するペプチドは高い結合親和性の値を有すると推測してもよく、細胞表面上に提示されないペプチドは低い結合親和性を有すると推測してもよい。例えば、高い結合親和性を有するペプチドはペプチドを細胞表面上に提示すると推測してもよく、低い結合親和性を有するペプチドはペプチドを細胞表面上に提示しないと推測してもよい。

【００５３】
一実施形態において、１つ以上のトレーニングデータセットのトレーニングデータに対して自己蒸留を実施してもよい。自己蒸留は、複数のサンプルに関する結合親和性及び提示の推定値を抽出することによって実施することができる。これらのサンプルは、上記トレーニングデータセットに対応するｗｅａｋ　ｌａｂｅｌに追加することができる。自己蒸留は、複対立分光法データを使用して実施することができる。自己蒸留は陽性の提示細胞を使用して実施することができる。未知の結合親和性を含むトレーニングデータセット中の陽性の提示細胞の場合、結合親和性は確立されたモデルを使用して推定することができる。

【００５４】
上記ニューラルネットワークモデルは、好ましくは、バイアスを制限するために、ポジティブトレーニングデータ及びネガティブトレーニングデータに対してトレーニングが行われる。不均衡なデータセットに対してネットワークのトレーニングを行うと、支配的なクラスのデータの表現をより多く学習することにより、上記ニューラルネットワークモデルにバイアスが生じる可能性があり、他のクラスが見落とされる可能性がある。例えば、ポジティブトレーニングデータセットのみに対してトレーニングを行ったニューラルネットワークモデルは、ペプチド腫瘍特異的ネオアンチゲンのＭＨＣクラスＩもしくはＭＨＣクラスＩＩ結合親和性を過大に予測する、または腫瘍特異的ネオアンチゲンがＭＨＣクラスＩもしくはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示されることになる確率を過大に予測する方向にバイアスが生じる可能性がある。ネガティブトレーニングセットのみに対してトレーニングを行ったニューラルネットワークモデルは、ペプチド腫瘍特

異的ネオアンチゲンのＭＨＣクラスＩもしくはＭＨＣクラスＩＩ結合親和性を過小に推定する、または腫瘍特異的ネオアンチゲンがＭＨＣクラスＩもしくはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示されることになる可能性を過小に推定する方向にバイアスが生じる可能性がある。

【００５５】

上記ＭＨＣクラスＩまたはＭＨＣクラスＩＩ親和性測定データセットは、通常、ポジティブトレーニングデータ及びネガティブトレーニングデータの両方を含む。例えば、ポジティブトレーニングデータとしては、陽性と分類される結合親和性予測値（例えば、＜５，０００ｎｍの結合親和性）を挙げることができる。例えば、ネガティブトレーニングデータとしては、陰性である結合親和性予測値（例えば、＞５，０００ｎｍの結合親和性）を挙げることができる。必要に応じて、親和性の低いランダムペプチドを含むようにトレーニングデータセットを拡張することにより、追加のネガティブトレーニングデータをトレーニングセットに組み込むことができる。例えば、上記ランダムペプチドは約＞２０，０００ｎｍの定性的な弱い親和性の目標を有していてもよい。

【００５６】

上記細胞表面へのペプチド提示トレーニングデータセットは、通常、ポジティブトレーニングデータ（例えば、ＭＨＣクラスＩタンパク質によって細胞表面上に提示されるペプチド）を含み、ネガティブトレーニングデータ（例えば、ＭＨＣクラスＩタンパク質によって細胞表面上に提示されることができないペプチド）は含まない。上記トレーニングデータセットがネガティブトレーニングデータを含まない場合には、ポジティブトレーニングデータセットを使用して、確率論的ネガティブトレーニングデータセット（例えば、ポジティブトレーニングデータセットから導出されるネガティブトレーニングデータセット）を生成させてもよい。ネガティブトレーニングデータセットは、ＨＬＡ対立遺伝子に対して「陽性」のペプチドをシャッフルすることによって生成させることができる。上記ペプチドは、アミノ酸長を変化させる（例えば、ペプチドをより長くまたはより短くする）ことによってシャッフルすることができる。あるいは、ペプチドのアミノ酸配列は、例えば、アミノ酸の置換、挿入、または欠失によって修飾することができる。挿入としては、アミノ末端及び／またはカルボキシル末端の融合、ならびに配列内挿入または複数のアミノ酸残基が挙げられる。欠失はペプチド配列からの１つ以上のアミノ酸残基の除去されることを特徴とする。アミノ酸置換は通常、単一の残基の置換であるが、複数の位置で生じる場合もある。置換、欠失、挿入、またはそれらの任意の組み合わせを組み合わせて、ＭＨＣクラスＩタンパク質またはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示されないペプチドに到達することができる。例えば、ＡＶＧＧＧＥＲＲＹＩＫＬのアミノ酸配列を有するペプチド配列を、ＣＶＧＧＧＥＨＲＹＩＭＮＮＬに修飾されてもよい。

【００５７】

加えて、またはペプチドシャッフリングと組み合わせて、ＨＬＡシャッフリングを使用し、ネガティブトレーニングデータセットを生成させることができる。「陽性」と分類されたＨＬＡ対立遺伝子（例えば、対応するペプチドを細胞表面上に提示するＨＬＡ対立遺伝子）が、陽性の対立遺伝子スーパータイプに属さない異なる対立遺伝子で置換されてもよい。

【００５８】

加えて、またはペプチドシャッフリング及び／またはＨＬＡシャッフリングと組み合わせて、ＨＲＶ陰性サンプリングを使用して、ネガティブトレーニングデータセットを生成させることができる。

【００５９】

上記トレーニングデータをさらにフィルタリングして、不必要なペプチドを除去してもよい。例えば、トレーニングデータセットが固有のペプチドを含むように、重複したペプチド（例えば、同一のアミノ酸配列）を除去してもよい。当業者であれば、ペプチドの同一性を判定する方法（すなわち、ペプチドが同一か異なるかを判定する方法）を容易に理解しよう。

【００６０】
トレーニングを行ったニューラルネットワークモデルを、免疫原性データセットを使用して検証してもよい。ニューラルネットワークモデルを検証することは、免疫原性データセットに１つ以上のランキング指標を適用することを含んでいてもよい。免疫原性検証データセット中のペプチドは、当該ペプチドの予測されるＭＨＣクラスＩまたはＭＨＣクラスＩＩ結合親和性と、該ペプチドがＭＨＣクラスＩまたはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示されることになる数値で表した確率とに基づいてランキングすることができる。上記ランキング指標はすべての対立遺伝子を集計してもよい。上記ランキング指標は、重み付けされた対立遺伝子頻度を使用することによって集計されてもよい。
【００６１】
実施形態において、上記ニューラルネットワークモデルを、ラベルなしのデータセットを使用してトレーニングを行ってもよい。例えば、細胞表面へのペプチド提示データセット中のペプチドはラベルなしであってもよい。理論に拘束されるものではないが、ラベルなしのデータセット（例えば、ペプチド配列）により、入力されたペプチド配列をより正確にキャラクタライズすることができる数ベクトル表現を提供すると考えられる。
【００６２】
ＩＩＩ．モデルのアーキテクチャ
本開示は、腫瘍特異的ネオアンチゲンのＭＨＣクラスＩまたはＭＨＣクラスＩＩ結合親和性と、各対象のペプチドに関する、対応するペプチドがＭＨＣクラスＩまたはＭＨＣクラスＩＩタンパク質によって細胞表面（すなわち、腫瘍細胞の表面）上に提示されることになる数値で表した確率とを共に予測するために、ニューラルネットワークモデルを使用することに関する。上記ニューラルネットワークモデルは、トレーニングにおいて該ニューラルネットワークモデルがこれまで遭遇したまたは遭遇したことのない腫瘍特異的ネオアンチゲンに好適である。
【００６３】
上記ニューラルネットワークモデルは、１つ以上の層中に配置された一連のノードを備える単一のニューラルネットワークであってよい。上記ノードは、それぞれが関連するパラメータを有する結合を介して、他のノードに結合されていてもよい。１つの特定のノードにおける値は、該特定のノードに関連付けられた活性化関数によってマッピングされた関連パラメータによって重み付けされた、特定のノードに結合したノードの値の合計として表わされてもよい。本明細書に記載の方法において使用されるニューラルネットワークモデルは、汎対立遺伝子モデル、対立遺伝子特異的モデル、スーパータイプ特異的モデル、またはそれらの組み合わせであってよい。
【００６４】
特定の一実施形態において、本方法は、腫瘍特異的ネオアンチゲンのペプチド配列を数値ベクトルに変換することを含む。通常、上記ペプチド配列は文字列（各文字がアミノ酸を表す）として表される。上記ペプチド配列は、当該ペプチドのアミノ酸及び該アミノ酸の位置を記述する情報を含む数ベクトルに変換することができる。上記数ベクトルは二項分類であってよい。例えば、ｋ^{i}個のアミノ酸を有するペプチド配列ｐ^{i}は、２０種のアミノ酸の行ベクトル（２０－ｋ）によって表され、該行ベクトルにおいて、当該ペプチド配列の特定の位置のアミノ酸のアルファベットに対応する単一の要素は１の値をもつことになる。残余の要素は０の値をもつことになる。例として、アミノ酸アルファベットがＡ、Ｃ、Ｄ、Ｅ、Ｆ、Ｇ、Ｈ、Ｉ、Ｋ、Ｌ、Ｍ、Ｎ、Ｐ、Ｑ、Ｒ、Ｓ、Ｔ、Ｖ、Ｗ、及びＹの場合、４個のアミノ酸のペプチド配列ＡＧＱＹは、８０個の要素の行ベクトル及び
【数８】

$$p^{i}=10000100000000100000110000100000000100000\ 11000010000000010000011000010000000010000001$$

によって表すことができる。

【００６５】
腫瘍特異的ネオアンチゲンのペプチド配列の長さは約５アミノ酸～約４０アミノ酸であってよい。例えば、上記ペプチド配列の長さは、５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０、２１、２２、２３、２４、２５、２６、２７、２８、２９、３０、３１、３２、３３、３４、３５、３６、３７、３８、３９、または４０アミノ酸であってよい。ＭＨＣクラスＩ分子は短鎖ペプチドに結合する。ＭＨＣクラスＩ分子は一般的に、長さが約５アミノ酸～約１０アミノ酸のペプチドを受け入れることができる。実施形態において、腫瘍特異的ネオアンチゲンのペプチド配列は、長さが約５アミノ酸～約１０アミノ酸の短鎖ペプチドである。ＭＨＣクラスＩＩ分子は長さがより長いペプチドに結合する。ＭＨＣクラスＩＩは一般的に、長さが約１３アミノ酸～約２５アミノ酸のペプチドを受け入れることができる。実施形態において、腫瘍特異的ネオアンチゲンのペプチド配列は、長さが約１３～２５アミノ酸の長いペプチドである。
【００６６】
腫瘍特異的ネオアンチゲンのペプチド配列の長さは同一であっても異なっていてもよい。腫瘍特異的ネオアンチゲンのペプチド配列が長さの異なるペプチド配列である（例えば、１つのペプチド配列の長さが７アミノ酸で、もう１つのペプチド配列の長さが１５アミノ酸である）場合には、それぞれの腫瘍特異的ネオアンチゲンのペプチドが隣接領域を含む最大のペプチド長さ（例えば、１５アミノ酸）に達するまで、パディング文字を当該数ベクトルに追加してもよい。パディング文字は隣接領域のＣ末端またはＮ末端に追加することができる。例として、上記パディング文字は、Ａ、Ｃ、Ｄ、Ｅ、Ｆ、Ｇ、Ｈ、Ｉ、Ｋ、Ｌ、Ｍ、Ｎ、Ｐ、Ｑ、Ｒ、Ｓ、Ｔ、Ｖ、Ｗ、及びＹのアミノ酸をコード化することになる。
【００６７】
上記腫瘍特異的ネオアンチゲンのペプチド配列は、該腫瘍特異的ネオアンチゲンペプチドに隣接する配列を含んでいてもよい。上記隣接配列は、上記腫瘍特異的ネオアンチゲンのペプチド配列のすぐ左側、上記腫瘍特異的ネオアンチゲンのペプチド配列のすぐ右側、またはその両方であってよい。
【００６８】
実施形態において、上記腫瘍特異的ネオアンチゲンのペプチド配列は、そのソースタンパク質配列内で該腫瘍特異的ネオアンチゲンペプチドに隣接するＣ末端配列の少なくとも１つ、またはそのソースタンパク質配列内で該腫瘍特異的ネオアンチゲンペプチドに隣接する少なくとも１つのＮ末端配列を含んでいてもよい。上記腫瘍特異的ネオアンチゲンのペプチド配列は、該腫瘍特異的ネオアンチゲンペプチドに隣接する少なくとも１つのＣ末端アミノ酸配列と、該腫瘍特異的ネオアンチゲンペプチドに隣接する少なくとも１つのアミノ酸Ｎ末端配列とを含むことが好ましい。
【００６９】
上記隣接領域の長さは、１、２、３、４、５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０、２１、２２、２３、２４、２５、２６、２７、２８、２９、３０、３１、３２、３３、３４、３５、３６、３７、３８、３９、４０アミノ酸、またはそれ以上であってよい。上記腫瘍特異的ネオアンチゲンペプチドのすぐ左側の隣接領域の長さは、１、２、３、４、５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０アミノ酸以上であってよい。上記腫瘍特異的ネオアンチゲンペプチドのすぐ右側の隣接領域の長さは、１、２、３、４、５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０アミノ酸以上であってよい。実施形態において、上記腫瘍特異的ネオアンチゲンのペプチド配列は、長さが最大約１０アミノ酸であってよい、該腫瘍特異的ネオアンチゲンのすぐ左側の隣接領域、及び／または長さが最大約１０アミノ酸であってよい、該腫瘍特異的ネオアンチゲンのすぐ右側の隣接領域を含む。上記隣接領域は、長さが５アミノ酸の、上記腫瘍特異的ネオアンチゲンのすぐ左側の隣接領域と、長さが５アミノ酸の、上記腫瘍特異的ネオアンチゲンのすぐ右側の隣接領域とを含むことが好ましい。上記隣接領域も同様に

、上記の数ベクトルにコード化されてもよい。

【００７０】

本方法は、ＨＬＡ対立遺伝子擬似配列を数ベクトルに変換することをさらに含む。上記ＨＬＡ対立遺伝子擬似配列はＨＬＡ対立遺伝子を表す。上記ＨＬＡ対立遺伝子擬似配列の長さは約５アミノ酸～約１００アミノ酸であってよい。例えば、上記ＨＬＡ対立遺伝子擬似配列の長さは、５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０、２１、２２、２３、２４、２５、２６、２７、２８、２９、３０、３１、３２、３３、３４、３５、３６、３７、３８、３９、４０、４１、４２、４３、４４、４５、４６、４７、４８、４９、５０、５１、５２、５３、５４、５５、５６、５７、５８、５９、６０、６１、６２、６３、６４、６５、６６、６７、６８、６９、７０、７１、７２、７３、７４、７５、７６、７７、７８、７９、８０、８１、８２、８３、８４、８５、８６、８７、８８、８９、９０、９１、９２、９３、９４、９５、９６、９７、９８、９９、または１００アミノ酸であってよい。上記ＨＬＡ対立遺伝子擬似配列の長さは約３０～約６０アミノ酸であってよい。本明細書に開示の方法の特定の実施形態において、上記ＨＬＡ対立遺伝子擬似配列の長さは約４０～約５０アミノ酸である。

【００７１】

上記ニューラルネットワークモデルの入力は、（ｉ）腫瘍特異的ネオアンチゲンのペプチド配列及び上記ペプチド配列の隣接領域を含む数ベクトル、及び（ｉｉ）ＨＬＡ対立遺伝子擬似配列を含む数ベクトルを含んでいてもよい。上記ニューラルネットワークの上記入力は、任意選択でセグメント識別子シーケンスを含んでいてもよい。セグメント識別子シーケンスは、各アミノ酸がいずれのセグメントに属するかを上記モデルに通知する。次に、腫瘍特異的ネオアンチゲンのペプチド配列及び上記ペプチド配列の隣接領域を含む数ベクトル、ならびに（ｉｉ）上記ＨＬＡ対立遺伝子擬似配列を含む数ベクトルが、１つ以上の埋め込み層にコード化される。上記埋め込み層は、腫瘍特異的ネオアンチゲンのペプチド配列及び上記のペプチド配列の隣接領域を含む高次元ベクトル、ならびに上記ＨＬＡ対立遺伝子擬似配列を含む数ベクトルを低次元空間に翻訳する。上記埋め込み層は、上記ニューラルネットワークモデルの最初の層と考えることができる。次いで、上記埋め込み層が平坦化されて、上記腫瘍特異的ネオアンチゲンの各ペプチド配列、上記ペプチド隣接領域、及びＨＬＡ対立遺伝子擬似配列の数ベクトル表現を生成させてもよい。

【００７２】

上記ペプチド腫瘍特異的ネオアンチゲンのＭＨＣクラスＩまたはＭＨＣクラスＩＩ結合親和性を予測するためには、上記腫瘍特異的ペプチド配列と上記ＨＬＡ対立遺伝子擬似配列とが連結される。これは、上記腫瘍特異的ネオアンチゲンのペプチド配列と上記ＨＬＡ対立遺伝子擬似配列とが鎖状または連続的に互いに結合していることを意味する。上記ＭＨＣクラスＩまたはＭＨＣクラスＩＩ結合親和性を予測するためには、上記隣接領域が連結される必要はない。必要ではないが、時には、上記隣接領域を連結することが望ましい場合もある。上記腫瘍特異的ネオアンチゲンのペプチド配列と上記ＨＬＡ擬似配列が連結されると、１つ以上のパラメータ（例えば、層及び／または関数）を適用することが可能になる。

【００７３】

適用することが可能な例示的な層としては、全結合ｄｅｎｓｅ　ｌａｙｅｒ、シーケンス層、活性化層、正規化層、ドロップアウト層、クロッピング層、プーリング層及びアンプーリング層、コンビネーション層（ｃｏｍｂｉｎａｔｉｏｎ　ｌａｙｅｒ）、物体検出層、または敵対的生成ネットワーク層が挙げられるが、これらに限定はされない。

【００７４】

全結合ｄｅｎｓｅ　ｌａｙｅｒの例としては、２Ｄ畳み込み層、３Ｄ畳み込み層、２Ｄグループ化畳み込み層、転置２Ｄ畳み込み層、転置３Ｄ畳み込み層、または全結合ｄｅｎｓｅ　ｌａｙｅｒが挙げられる。シーケンス層の例としては、シーケンス入力層、ＬＳＴＭ層、双方向ＬＳＴＭ層、ＧＲＵ層、シーケンス折り畳み層、シーケンス展開層、平坦化層、または単語埋め込み層が挙げられる。活性化層の例としては、ＲｅＬＵ層、ｌｅａｋ

ｙ　ＲｅＬＵ層、クリップされたＲｅＬＵ層、ＥＬＵ活性化層、双曲線正接活性化層、またはＰＲｅＬＵ層が挙げられる。正規化層、ドロップアウト層、クロッピング層の例としては、バッチ正規化層、グループ正規化層、チャネル単位の局所応答正規化層、ドロップアウト層、２Ｄクロップ層、３Ｄクロップ層、２Ｄリサイズ層、３Ｄリサイズ層が挙げられる。プーリング層及びアンプーリング層の例としては、平均プーリング層、３Ｄ層、グローバル平均プーリング層、３Ｄグローバル平均プーリング層、最大値プーリング層、３Ｄ最大値プーリング層、グローバル最大値プーリング層または最大値アンプーリング層が挙げられる。コンビネーション層層の例としては、加算層、乗算層、深さ連結層、加重平均層が挙げられる。物体検出層の例としては、ＲＯＩ入力層、ＲＯＩ最大値プーリング層、ＲＯＩ　ａｌｉｇｎ層、アンカーボックス層、領域提案層、ＳＳＤマージ層、空間から深さへの変換層、領域提案ネットワーク、Ｆｏｃａｌ　Ｌｏｓｓ層、領域提案ネットワーク、ボックス回帰が挙げられる。

【００７５】

実施形態において、１つ以上の全結合ｄｅｎｓｅ　ｌａｙｅｒが適用されてもよい。全結合層は、入力に重み行列を乗じ、次いでバイアスベクトルを追加する。例えば、１、２、３、４、５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０、またはそれ以上の全結合ｄｅｎｓｅ　ｌａｙｅｒを適用することができる。ＭＨＣクラスＩまたはＭＨＣクラスＩＩ結合親和性を予測する場合、少なくとも３つの全結合ｄｅｎｓｅ　ｌａｙｅｒが適用されることが好ましい。

【００７６】

実施形態において、１つ以上の活性化層（関数）が適用されてもよい。上記活性化関数は、ニューロンまたは該ニューロンの層全体に割り当てることができる。適用することができる例示的な活性化関数は、ＥＬＵ活性化関数またはｒｅＬＵ層である。上述の及び／または当業者に公知の他の活性化層が適用されてもよい。上記活性化関数は、合算された、重み付けされたノードからの入力を、ノードまたは出力の活性化に変換することができる。例えば、１、２、３、４、５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０、またはそれ以上の活性化層（関数）が適用されてもよい。通常、約１、２、３、４、または５つの活性化関数が適用されてもよい。

【００７７】

１つ以上のドロップアウト層が適用されてもよい。ドロップアウト層は過学習を低減するのに有利であり、それによってよりよい結果が得られる。例えば、１、２、３、４、５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０、またはそれ以上のドロップアウト層が適用されてもよい。通常、約１、２、３、４、または５つのドロップアウト層が適用されてもよい。

【００７８】

例示的なニューラルネットワークモデルにおいて、１つ以上の全結合ｄｅｎｓｅ　ｌａｙｅｒ、１つ以上の活性化関数、及び１つ以上のドロップアウト層が適用されてもよい。好ましいニューラルネットワークモデルにおいて、１つ以上の全結合ｄｅｎｓｅ　ｌａｙｅｒ、活性化関数（例えば、ＥＬＵ活性化関数）、及び１つ以上のドロップアウト層が適用されてもよい。

【００７９】

シーケンス表現のより良い学習を可能にするために、１つ以上のＬＳＴＭ層または１つ以上の双方向ＬＴＳＭ層が適用されてもよい。Ｔｒａｎｓｆｏｒｍｅｒも追加してよい。例えば、約１、約２、約３、約４、約５、約６、約７、約８、約９、約１０、約１１、約１２、約１３、約１４、約１５、またはそれ以上のＴｒａｎｓｆｏｒｍｅｒ層が追加されてもよい。Ｔｒａｎｓｆｏｒｍｅｒは埋め込まれたアミノ酸配列中に位置の埋め込みを追加することが可能であり、１つ以上の積層エンコーダー層を含んでいてもよい。例えば、Ｔｒａｎｓｆｏｒｍｅｒは、１つ以上のマルチヘッドアテンション層、１つ以上のドロップアウト層、１つ以上の正規化層、１つ以上のフィードフォワード層、またはそれらの組み合わせを含んでいてもよい。例示的なＴｒａｎｓｆｏｒｍｅｒは、（１）マルチヘッド

アテンション層、（２）ドロップアウト層（０．１の比）、（３）正規化層、（４）フィードフォワード層（線形層及びＲｅＬＵ層）、（５）ドロップアウト層（０．１の比）、及び（６）層正規化を含んでいてもよい。

【００８０】

回帰モデルを適用して、上記ペプチド腫瘍特異的ネオアンチゲンのＭＨＣクラスＩまたはクラスＩＩ結合親和性を予測することができる。特に、平均二乗損失関数の変化形を使用することができる。平平均二乗損失関数はL_{BA-MSE}によって表すことができ、但し、測定値は（＞）または（＜）を伴い、データセット中の定量的及び定性的なペプチド－ＭＨＣ結合親和性測定値の両方の処理に関して、上記測定値は、不等式に反する場合にのみ損失に寄与する。

【００８１】

上記出力はペプチドリガンド－ＭＨＣクラスＩまたはＭＨＣクラスＩＩの結合親和性を表す数値スコアを含む。

【００８２】

上記ニューラルネットワークモデルは、上記腫瘍特異的ネオアンチゲンがＭＨＣクラスＩまたはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示されることになる確率を共に予測することを含む。上記腫瘍特異的ネオアンチゲンがＭＨＣクラスＩタンパク質によって細胞表面上に提示されることになる確率を予測するために、上記腫瘍特異的ペプチド配列、対応する隣接領域、及びＨＬＡ対立遺伝子擬似配列が、単一の数値スコアに連結される。

【００８３】

上記腫瘍特異的ネオアンチゲンのペプチド配列、上記対応する隣接領域、及び上記ＨＬＡ擬似配列が連結されたところで、１つ以上のパラメータ（例えば、層及び／または関数）が適用されてもよい。

【００８４】

１つ以上の全結合ｄｅｎｓｅ　ｌａｙｅｒが適用されてもよい。例えば、１、２、３、４、５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０、またはそれ以上の全結合ｄｅｎｓｅ　ｌａｙｅｒが適用されてもよい。上記腫瘍特異的ネオアンチゲンがＭＨＣクラスＩまたはクラスＩＩタンパク質によって細胞表面上に提示されることになる確率を予測するためには、少なくとも３つの全結合ｄｅｎｓｅ　ｌａｙｅｒが適用されることが好ましい。

【００８５】

１つ以上の活性化関数がニューロンまたは該ニューロンの層全体に割り当てられてもよい。適用することができる例示的な活性化関数は、ＥＬＵ活性化関数またはｒｅＬＵ層である。上述の及び／または当業者に公知の他の活性化層が適用されてもよい。上記活性化関数は、合算された、重み付けされたノードからの入力を、ノードまたは出力の活性化に変換することができる。例えば、１、２、３、４、５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０、またはそれ以上の活性化層（関数）が適用されてもよい。通常、約１、２、３、４、または５つの活性化関数が適用されてもよい。

【００８６】

１つ以上のドロップアウト層が適用されてもよい。ドロップアウト層は過学習を低減するのに有利であり、それによってよりよい結果が得られる。例えば、１、２、３、４、５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０、またはそれ以上のドロップアウト層が適用されてもよい。通常、約１、２、３、４、または５つのドロップアウト層が適用されてもよい。

【００８７】

Ｆｏｃａｌ　Ｌｏｓｓ二値分類を使用して、（上記のスキーム３を参照のこと）を使用して、上記腫瘍特異的ネオアンチゲンがＭＨＣクラスＩまたはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示されることになる確率を予測することができる。上記出力は

、ペプチドがＭＨＣクラスＩまたはＭＨＣクラスＩＩタンパク質によって細胞表面上に提示されることになる数値で表した確率である。

【００８８】

上記ニューラルネットワークモデルに、さらにキャリブレーションが行われてもよい。モデルにキャリブレーションが行われていない場合、ニューラルネットワークは確率を過大に予測するまたは過小に予測する可能性がある。したがって、本明細書に記載のニューラルネットワークにキャリブレーションを行うことにより、予測される確率の精度及び信頼性を向上させることができる。上記ニューラルネットワークにキャリブレーションを行うために、確率計算が適用されてもよい。特に、確率計算が１つ以上の対象のＨＬＡ対立遺伝子に適用されてもよい。例えば、１、２、３、４、５、または６個のＨＬＡ対立遺伝子。確率計算を使用して、各ＨＬＡ対立遺伝子に対するモデルの予測に基づいて、対象の対立遺伝子の全体的な提示の確率を推測することができる。検証データセットに対してニューラルネットワークの提示の予測にキャリブレーションを行うことによって、上記ニューラルネットワークにさらにキャリブレーションを行うことができる。例えば、キャリブレーション曲線に低次の多項式が適用されてもよい。多項式係数を正に制約して、単調増加関数を得することができる。Ｌａｓｓｏ線形回帰が使用されてもよい。キャリブレーションを行った提示の予測（例えば、上記腫瘍特異的ネオアンチゲンのＭＨＣクラスＩ結合親和性及び上記腫瘍特異的ネオアンチゲンがＭＨＣクラスＩタンパク質によって細胞表面上に提示される確率）は、免疫原性の代用値として使用することができる。

【００８９】

上記ニューラルネットワークモデルのパフォーマンスを、本明細書に記載の免疫原性データを使用して評価することができる。上記ニューラルネットワークの予測は、１つ以上のランキング指標を使用して評価することができる。例示的なランキング指標としては、上位ｋ位の項目、Ｐｒｅｃｉｓｉｏｎ@Ｋ、$_{n}$ＤＣＧ$_{K}$、逆順位、及び陽性的中率指標が挙げられるが、これらに限定はされない。免疫原性データセット中の各対立遺伝子に対応するすべてのペプチドは、１つ以上のランキング指標を使用して、予測される細胞表面へのペプチド提示の確率及び／または予測される結合親和性スコアに基づいてランキングすることができる。ついで、重み付けされた対立遺伝子頻度を使用してランキング指標を集計することができる。

【００９０】

一実施形態において、自己監視型事前トレーニングが実施されてもよい。例示的なトレーニングモデルとしては、マスクされた言語モデリング及び次のペプチド予測がある。

【００９１】

ＩＶ．コンピュータによる方法の実施

プログラムされた、または他の形態で構成されたコンピュータシステムを使用して、本明細書に開示の方法を実施することができる。上記コンピュータシステムは、単一の計算装置、または１つ以上の計算ネットワークを使用して相互接続された複数の計算装置を備えていてもよい。上記コンピュータシステムは、コンピュータの能力を使用して、本明細書に記載のニューラルネットワークモデルを実行することができる。

【００９２】

上記コンピュータシステムは中央演算装置を備えていてもよく、上記装置は、シングルコアもしくはマルチコアプロセッサ、または並列処理用の複数のプロセッサであってよい。上記システムは、メモリ（ランダムアクセスメモリ、読取り専用メモリ、フラッシュメモリなど）、電子記憶装置（例えば、クラウドプラットフォーム）、１つ以上のシステムと通信するための通信インターフェース、ならびにデータ記憶装置、その他のメモリ、及びディスプレイアダプターなどの他の周辺装置を備えていてもよい。上記メモリ、記憶装置、インターフェース、及び周辺装置は、通信バスを介してＣＰＵと通信していてもよい。これらのコンポーネントのいずれか１つまたはすべては、共有内部ネットワークまたは外部ネットワークを介して通信していてもよく、且つ集合システムは上記ネットワークを介して１つ以上のユーザーデバイスと通信していてもよい。上記ネットワークは、インタ

ーネット、エクストラネット、またはインターネットと通信しているインターネット／エクストラネットであってよい。上記ネットワークは１つ以上のコンピュータサーバーを備えていてもよく、該コンピュータサーバーは、クラウドコンピューティングなどの分散計算を可能にするものであってよい。上記コンピュータシステムは処理システムと通信していてもよい。上記処理システムは、本明細書に開示の方法を実施するように構成されていてもよい。

【００９３】

上記計算装置のさまざまな例としては、デスクトップコンピュータ、ラップトップ、及び携帯電話、タブレットコンピュータ、パーソナルコンピュータ、ウェアラブルコンピュータ、サーバー、携帯情報端末（ＰＤＡ）、ハイブリッドＰＤＡ／携帯電話、携帯電話、電子ブックリーダー、セットトップボックス、音声コマンドデバイス、カメラ、デジタルメディアプレーヤーなどが挙げられるが、これらに限定はされない。いくつかの実施形態において、上記コンピュータデバイスは、トレーニングの要求、展開の要求、及び／または実行の要求を送信するための、１つ以上のユーザインターフェース、コマンドラインインターフェース（ＣＬＩ）、アプリケーションプログラミングインターフェース（ＡＰＩ）、及び／または他のプログラムインターフェースを有していてもよい。いくつかの実施形態において、実施形態において、上記コンピュータデバイスは、上記ニューラルネットワークモデルと対話するスタンドアロンアプリケーションを実行してもよい。

【００９４】

いくつかの実施形態において、上記ネットワークとしては、任意の有線ネットワーク、無線ネットワーク、またはそれらの組み合わせが挙げられる。例えば、上記ネットワークは、パーソナルエリアネットワーク、ローカルエリアネットワーク、広域ネットワーク、無線ブロードキャストネットワーク（例えば、ラジオまたはテレビ用）、ケーブルネットワーク、衛星ネットワーク、携帯電話ネットワーク、またはそれらの組み合わせであってよい。さらなる例として、上記ネットワークは、場合によりインターネットなどのさまざまな別個の当事者によって運営されている、公的にアクセス可能な連結したネットワークのネットワークであってもよい。いくつかの実施形態において、上記ネットワークは、企業または大学のイントラネットなどのプライベートまたは半プライベートネットワークであってもよい。上記ネットワークは、Ｇｌｏｂａｌ Ｓｙｓｔｅｍ ｆｏｒ Ｍｏｂｉｌｅ Ｃｏｍｍｕｎｉｃａｔｉｏｎｓ （ＧＳＭ）ネットワーク、Ｃｏｄｅ Ｄｉｖｉｓｉｏｎ Ｍｕｌｔｉｐｌｅ Ａｃｃｅｓｓ （ＣＤＭＡ）ネットワーク、Ｌｏｎｇ Ｔｅｒｍ Ｅｖｏｌｕｔｉｏｎ （ＬＴＥ） ネットワーク、または任意の他の形態の無線ネットワークなどの、１つ以上の無線ネットワークを含んでいてもよい。上記ネットワークは、インターネットまたは上述の他の形態のネットワークのいずれかを介して通信するためのプロトコル及びコンポーネントを使用してもよい。例えば、上記ネットワークによって使用されるプロトコルとしては、ＨＴＴＰ、ＨＴＴＰ Ｓｅｃｕｒｅ（ＨＴＴＰＳ）、Ｍｅｓｓａｇｅ Ｑｕｅｕｅ Ｔｅｌｅｍｅｔｒｙ Ｔｒａｎｓｐｏｒｔ （ＭＱＴＴ）、Ｃｏｎｓｔｒａｉｎｅｄ Ａｐｐｌｉｃａｔｉｏｎ Ｐｒｏｔｏｃｏｌ （ＣｏＡＰ）などを挙げることができる。インターネットまたはその他の上述の形態の通信ネットワークのいずれかを介して通信するためのプロトコル及びコンポーネントは当業者には周知であり、よって本明細書ではこれ以上詳細な説明は行わない。

【００９５】

図９は、いくつかの実施形態に係るプロバイダネットワーク（または「サービスプロバイダシステム」）環境の例を示す。プロバイダネットワーク９００は、顧客が、プロバイダネットワーク内または１つ以上のデータセンターにおけるネットワーク内のデバイス上で実行される計算及びストレージリソースを含む、但しこれらに限定されない仮想化リソースのインスタンス９１２を、購入、レンタル、または他の形態で取得することを可能にする、１つ以上の仮想化サービス９１０を介して顧客にリソース仮想化を提供してもよい。ローカルインターネットプロトコル（ＩＰ）アドレス９１６は、リソースインスタンス９１２に関連付けられてもよく、ローカルＩＰアドレスはプロバイダネットワーク９００

上のリソースインスタンス９１２の内部ネットワークアドレスである。いくつかの実施形態において、プロバイダネットワーク９００は、顧客がプロバイダ９００から取得することができるパブリックＩＰアドレス９１４及び／またはパブリックＩＰアドレス範囲（例えば、インターネットプロトコルバージョン４（ＩＰｖ４）またはインターネットプロトコルバージョン６（ＩＰｖ６）アドレス）の提供も行ってよい。

【００９６】

従来、プロバイダネットワーク９００により、仮想化サービス９１０を介して、サービスプロバイダの顧客（例えば、１つ以上の顧客デバイス９５２を含む１つ以上のクライアントネットワーク９５０Ａ～９５０Ｃを運用する顧客）が、当該顧客に割り当てられた（ａｓｓｉｇｎｅｄ）かまたは割り振られた（ａｌｌｏｃａｔｅｄ）少なくともいくつかのパブリックＩＰアドレス９１４を、該顧客に割り当てられた特定のリソースインスタンス９１２に動的に関連付けることが可能になり得る。プロバイダネットワーク９００により、上記顧客が、該顧客に割り振られた１つの仮想化計算リソースインスタンス９１２に以前にマッピングされたパブリックＩＰアドレス９１４を、同様に該顧客に割り振られた別の仮想化計算リソースインスタンス９１２に再マッピングすることも可能になり得る。サービスプロバイダによって提供される仮想化計算リソースインスタンス９１２及びパブリックＩＰアドレス９１４を使用して、顧客ネットワーク（複数可）９５０Ａ～９５０Ｃのオペレータなどのサービスプロバイダの顧客は、例えば、顧客固有のアプリケーションを実行し、インターネットなどの中間ネットワーク９４０上の顧客のアプリケーションを提供してもよい。次いで、中間ネットワーク９４０上のその他のネットワークエンティティ９２０が、顧客ネットワーク（複数可）９５０Ａ～９５０Ｃによって公開される宛先パブリックＩＰアドレス９１４へのトラフィックを生成してもよく、該トラフィックはサービスプロバイダのデータセンターにルーティングされ、該データセンターでは、ネットワーク基板を介して、現時点で宛先パブリックＩＰアドレス９１４にマッピングされている仮想化計算リソースインスタンス９１２のローカルＩＰアドレス９１６にルーティングされる。同様に、仮想化計算リソースインスタンス９１２からの応答トラフィックは、ネットワーク基板を介して、中間ネットワーク９４０上のソースエンティティ９２０に戻されてもよい。

【００９７】

本明細書では、ローカルＩＰアドレスとは、例えばプロバイダネットワーク中のリソースインスタンスの内部または「プライベート」ネットワークアドレスを指す。ローカルＩＰアドレスは、Ｉｎｔｅｒｎｅｔ Ｅｎｇｉｎｅｅｒｉｎｇ Ｔａｓｋ Ｆｏｒｃｅ（ＩＥＴＦ） Ｒｅｑｕｅｓｔ ｆｏｒ Ｃｏｍｍｅｎｔｓ（ＲＦＣ） １９１８によって予約されたアドレスブロック内、及び／またはＩＥＴＦ ＲＦＣ ４１９３によって指定されたアドレス形式のアドレスブロック内に存在してもよく、プロバイダネットワーク内で変更可能であってもよい。プロバイダネットワークの外部起源のネットワークトラフィックは、ローカルＩＰアドレスに直接ルーティングされず、代わりに、該トラフィックはリソースインスタンスのローカルＩＰアドレスにマッピングされたパブリックＩＰアドレスを使用する。プロバイダネットワークは、パブリックＩＰアドレスからローカルＩＰアドレスへの、及びその逆のマッピングを実行するネットワークアドレス変換（ＮＡＴ）もしくは類似の機能を提供するネットワーキングデバイスまたはアプライアンスを備えていてもよい。

【００９８】

パブリックＩＰアドレスは、サービスプロバイダまたは顧客のいずれかによってリソースインスタンスに割り当てられる、インターネットの変更可能なネットワークアドレスである。パブリックＩＰアドレスにルーティングされたトラフィックは、例えば１：１ ＮＡＴを介して翻訳され、リソースインスタンスのそれぞれのローカルＩＰアドレスに転送される。

【００９９】

いくつかのパブリックＩＰアドレスは、プロバイダーネットワークインフラストラクチ

ャによって特定のリソースインスタンスに割り当てられてもよく、これらのパブリックＩＰアドレスは、標準パブリックＩＰアドレス、または単に標準ＩＰアドレスと呼ばれる場合がある。いくつかの実施形態において、標準ＩＰアドレスのリソースインスタンスのローカルＩＰアドレスへのマッピングは、すべてのリソースインスタンスの形態のデフォルトの起動構成である。

【０１００】

少なくともいくつかのパブリックＩＰアドレスは、プロバイダネットワーク９００の顧客に割り振られるか、またはプロバイダネットワーク９００の顧客によって取得されてもよく、次いで顧客は、割り振られたパブリックＩＰアドレスを、当該顧客に割り振られた特定のリソースインスタンスに割り振ってもよい。これらのパブリックＩＰアドレスは、顧客のパブリックＩＰアドレス、または単に顧客のＩＰアドレスと呼ばれる場合がある。顧客ＩＰアドレスは、標準ＩＰアドレスの場合のようにプロバイダネットワーク９００によってリソースインスタンスに割り当てられる代わりに、例えばサービスプロバイダによって提供されるＡＰＩを介して、当該顧客によってリソースインスタンスに割り当てられてもよい。標準ＩＰアドレスとは異なり、顧客ＩＰアドレスは顧客アカウントに割り振られ、必要または希望に応じてそれぞれの顧客によって他のリソースインスタンスに再マッピングされてもよい。顧客ＩＰアドレスは、特定のリソースインスタンスではなく、顧客のアカウントに関連付けられており、当該顧客がそのＩＰアドレスを放棄することを選択するまで、該顧客が該ＩＰアドレスを管理する。顧客ＩＰアドレスにより、従来の静的ＩＰアドレスとは異なり、顧客のパブリックＩＰアドレスを顧客のアカウントに関連付けられたいずれかのリソースインスタンスに再マッピングすることで、顧客がリソースインスタンスまたはアベイラビリティゾーンの障害をマスクすることが可能になる。例えば、顧客ＩＰアドレスにより、顧客のＩＰアドレスを代替リソースインスタンスに再マッピングすることで、顧客が、顧客のリソースインスタンスまたはソフトウェアの問題を回避することが可能になる。

【０１０１】

図１０は、いくつかの実施形態に係る、ストレージサービス及びハードウェア仮想化サービスを顧客に提供する例示的なプロバイダネットワークのブロック図である。ハードウェア仮想化サービス１０２０は、複数の計算資源１０２４（例えば、ＶＭ）を顧客に提供する。計算資源１０２４は、例えば、プロバイダネットワーク１０００の顧客（例えば、顧客ネットワーク１０５０を実行する顧客）にレンタルまたはリースされてもよい。各計算資源１０２４には、１つ以上のローカルＩＰアドレスが与えられてもよい。プロバイダネットワーク１０００は、計算資源１０２４のローカルＩＰアドレスからパブリックインターネットの宛先へ、及びパブリックインターネットソースから計算資源１０２４のローカルＩＰアドレスへパケットをルーティングするように構成されていてもよい。

【０１０２】

プロバイダネットワーク１０００は、例えばローカルネットワーク１０５６を介して中間ネットワーク１０４０に連結した顧客ネットワーク１０５０、ハードウェア仮想化サービス１０２０を介して中間ネットワーク１０４０及びプロバイダネットワーク１０００に連結した仮想計算システム１０９２を実行する能力を提供してもよい。いくつかの実施形態において、ハードウェア仮想化サービス１０２０は、１つ以上のＡＰＩ　１００２、例えばウェブサービスインターフェース、を提供してもよく、これを介して、顧客ネットワーク１０５０は、例えばコンソール１０９４（例えば、ウェブ上のアプリケーション、スタンドアロンのアプリケーション、モバイルのアプリケーションなど）を介し、ハードウェア仮想化サービス１０２０によって提供される機能にアクセスすることができる。いくつかの実施形態において、プロバイダネットワーク１０００では、顧客ネットワーク１０５０の各仮想計算システム１０９２が、顧客ネットワーク１０５０にリースされ、レンタルされ、またはその他の形態で提供される計算リソース１０２４に対応することができる。

【０１０３】

仮想計算システム１０９２及び／または別の顧客デバイス１０９０のインスタンスから（例えば、コンソール１０９４を介して）、顧客が、例えば１つ以上のＡＰＩ１００２を介してストレージサービス１０１０の機能にアクセスして、プロバイダネットワーク１０００によって提供される仮想データストア１０１６のストレージリソース１０１８Ａ～１０１８Ｎ（例えば、フォルダまたは「バケット」、仮想化ボリューム、データベースなど）のデータにアクセスする、及びここにデータを格納することができる。いくつかの実施形態において、仮想化データストアゲートウェイ（図示せず）が顧客ネットワーク１０５０に提供されてもよく、該ゲートウェイは、少なくとも一部のデータ、例えば、頻繁にアクセスされるデータまたは重要なデータをローカルにキャッシュしてもよく、且つ、データの一次ストア（仮想化データストア１０１６）が維持されるように、ローカルキャッシュから新しいデータまたは変更されたデータをアップロードするための１つ以上の通信チャネルを介して、ストレージサービス１０１０と通信してもよい。いくつかの実施形態において、ユーザは、仮想計算システム１０９２を介して及び／または別の顧客デバイス１０９０上で、仮想データストア１０１６のボリュームをマウントし、且つストレージ仮想化サービスとして機能するストレージサービス１０１０を介して上記１０１６のボリュームにアクセスすることができ、これらのボリュームは、上記ユーザにはローカル（仮想化）ストレージ１０９８として見える場合がある。

【０１０４】

図１０には示されていないが、仮想化サービス（複数可）は、ＡＰＩ（複数可）１００２を介してプロバイダネットワーク１０００内のリソースインスタンスからアクセスすることもできる。例えば、顧客、機器サービスプロバイダ、または他のエンティティは、プロバイダネットワーク１０００上のそれぞれの仮想ネットワーク内からＡＰＩ　１００２を介して仮想化サービスにアクセスし、上記仮想ネットワーク内または別の仮想ネットワーク内の１つ以上のリソースインスタンスの割り振りを要求することができる。

【０１０５】

いくつかの実施形態において、本明細書に記載の技法の一部またはすべてを実施するシステムは、図１１に示すコンピュータシステム１１００などの、コンピュータにアクセス可能な１つ以上の媒体を備える、または上記媒体にアクセスするように構成された汎用コンピュータシステムを備えていてもよい。図示した実施形態において、コンピュータシステム１１００は、入出力（Ｉ／Ｏ）インターフェース１１３０を介してシステムメモリ１１２０に連結した１つ以上のプロセッサ１１１０を備える。コンピュータシステム１１００は、Ｉ／Ｏインターフェース１１３０に連結したネットワークインターフェース１１４０をさらに備える。図１１は単一の計算デバイスとしてのコンピュータシステム１１００を示しているが、さまざまな実施形態において、コンピュータシステム１１００は、１つの計算デバイス、または単一のコンピュータシステム１１００として協働するように構成された任意の数の計算デバイスを備えていてもよい。

【０１０６】

さまざまな実施形態において、コンピュータシステム１１００は、１つのプロセッサ１１１０を備えるユニプロセッサシステムであっても、またはいくつかのプロセッサ１１１０（例えば、２つ、４つ、８つ、もしくは別の適切な数）を備えるマルチプロセッサシステムであってもよい。プロセッサ１１１０は、命令を実行する能力をもつ任意且つ適宜のプロセッサであってよい。例えば、さまざまな実施形態において、プロセッサ１１１０は、さまざまな命令セットアーキテクチャ（ＩＳＡ）、例えば、ｘ８６、ＡＲＭ、ＰｏｗｅｒＰＣ、ＳＰＡＲＣ、もしくはＭＩＰＳ　ＩＳＡ、または任意且つ適宜の他のＩＳＡなどのいずれかを実装する汎用プロセッサあるいは埋め込みプロセッサであってよい。マルチプロセッサシステムにおいては、プロセッサ１１１０のそれぞれは、一般的に同一のＩＳＡを実装してよいが、必ずしもそうである必要はない。

【０１０７】

システムメモリ１１２０は、プロセッサ（複数可）１１１０によってアクセス可能な命令及びデータを格納することができる。さまざまな実施形態において、システムメモリ１

１２０は、ランダムアクセスメモリ（ＲＡＭ）、スタティックＲＡＭ（ＳＲＡＭ）、同期ダイナミックＲＡＭ（ＳＤＲＡＭ）、不揮発性／フラッシュタイプメモリ、または任意の他の種類のメモリなどの、任意且つ適宜のメモリ技術を使用して実装することができる。図示した実施形態において、上述の方法、技法、及びデータなどの、１つ以上の所望の機能を実行するプログラム命令及びデータが、システムメモリ１１２０内に格納されることが示されている。

【０１０８】

一実施形態において、Ｉ／Ｏインターフェース１１３０は、プロセッサ１１１０、システムメモリ１１２０、及びネットワークインターフェース１１４０または他の周辺インターフェースを含む、デバイス中の任意の周辺デバイスの間のＩ／Ｏトラフィックを調整するように構成されてもよい。いくつかの実施形態において、Ｉ／Ｏインターフェース１１３０は、１つのコンポーネント（例えば、システムメモリ１１２０）からのデータ信号を別のコンポーネント（例えば、プロセッサ１１１０）による使用に適したフォーマットへと変換するために、任意の必要なプロトコル、タイミング、または他のデータ変換を実行してもよい。いくつかの実施形態において、Ｉ／Ｏインターフェース１１３０は、例えば、Ｐｅｒｉｐｈｅｒａｌ Ｃｏｍｐｏｎｅｎｔ Ｉｎｔｅｒｃｏｎｎｅｃｔ （ＰＣＩ）バス規格またはＵｎｉｖｅｒｓａｌ Ｓｅｒｉａｌ Ｂｕｓ （ＵＳＢ）規格の変化形などの、さまざまな種類の周辺バスを介して接続されたデバイスのサポートを含んでいてもよい。いくつかの実施形態において、Ｉ／Ｏインターフェース１１３０の機能は、例えばノースブリッジ及びサウスブリッジなどの２つ以上の別個のコンポーネントに分割されてもよい。また、いくつかの実施形態において、システムメモリ１１２０へのインターフェースなどのＩ／Ｏインターフェース１１３０の機能の一部またはすべてが、プロセッサ１１１０中に直接組み込まれていてもよい。

【０１０９】

ネットワークインターフェース１１４０は、コンピュータシステム１１００と、例えば図１に示す他のコンピュータシステムもしくはデバイスなどのネットワークまたは複数のネットワーク１１５０に接続された他のデバイス１１６０との間でデータを交換することが可能になるように構成されていてもよい。さまざまな実施形態において、ネットワークインターフェース１１４０は、例えばＥｔｈｅｒｎｅｔネットワークの形態などの、任意且つ適宜の有線または無線の一般データネットワークを介した通信をサポートすることができる。さらに、ネットワークインターフェース１１４０は、アナログ音声ネットワークまたはデジタルファイバ通信ネットワークなどの遠隔通信／電話ネットワークを介した、Ｆｉｂｅｒ Ｃｈａｎｎｅｌ ＳＡＮなどのストレージエリアネットワーク（ＳＡＮ）を介した、もしくはＩ／Ｏの任意且つ適宜の他の形態のネットワーク及び／またはプロトコルを介した通信をサポートしてもよい。

【０１１０】

いくつかの実施形態において、コンピュータシステム１１００は、Ｉ／Ｏインターフェース１１３０（例えば、Ｐｅｒｉｐｈｅｒａｌ Ｃｏｍｐｏｎｅｎｔ Ｉｎｔｅｒｃｏｎｎｅｃｔ Ｅｘｐｒｅｓｓ（ＰＣＩ－Ｅ）標準のバージョン、またはＱｕｉｃｋＰａｔｈインターコネクト（ＱＰＩ）もしくはＵｌｔｒａＰａｔｈインターコネクト（ＵＰＩ）などの別のインターコネクトを実装するバス）を使用して結合した１つ以上のオフロードカード１１７０（１つ以上のプロセッサ１１７５を含み、場合によっては１つ以上のネットワークインターフェース１１４０を含む）を含む。例えば、いくつかの実施形態において、コンピュータシステム１１００は、計算インスタンスをホストするホスト電子デバイス（例えば、ハードウェア仮想化サービスの一部として動作する）として機能してもよく、１つ以上のオフロードカード１１７０は、ホスト電子デバイス上で実行される計算インスタンスを管理することができる仮想化マネージャを実行する。一例として、いくつかの実施形態において、オフロードカード（複数可）１１７０は、計算インスタンスの一時停止及び／または一時停止解除、計算インスタンスの起動及び／または終了、メモリ転送／コピー操作の実行などの計算インスタンス管理操作を実行してもよい。これらの管理操作は

、いくつかの実施形態において、コンピュータシステム１１００の他のプロセッサ１１１０Ａ～１１１０Ｎによって実行されるハイパーバイザと連携して（例えば、ハイパーバイザからの要求に応じて）オフロードカード（複数可）１１７０によって実行されてもよい。しかしながら、いくつかの実施形態において、）オフロードカード（複数可）１１７０によって実装される仮想化マネージャは、他のエンティティからの（例えば、計算インスタンス自体からの）要求に対応することができ、任意の別個のハイパーバイザと連携（またはサービス）しない場合がある。

【０１１１】

いくつかの実施形態において、システムメモリ１１２０は、上述のようにプログラム命令及びデータを記憶するように構成されたコンピュータアクセス可能な媒体の一実施形態であってもよい。しかしながら、他の実施形態において、プログラム命令及び／またはデータは、受信され、送信され、またはさまざまな種類のコンピュータがアクセス可能な媒体上に格納されてもよい。一般的に言えば、コンピュータアクセス可能な媒体としては、Ｉ／Ｏインターフェース１１３０を介してコンピュータシステム１１００に連結した磁気媒体もしくは光媒体、例えば、ディスクもしくはＤＶＤ／ＣＤなどの非一時的記憶媒体または記憶媒体を挙げることができる。非一時的なコンピュータアクセス可能な記憶媒体としては、ＲＡＭ（例えば、ＳＤＲＡＭ、ダブルデータレート（ＤＤＲ）ＳＤＲＡＭ、ＳＲＡＭなど）、読み取り専用メモリ（ＲＯＭ）などの任意の揮発性または不揮発性媒体も挙げることができ、これは、システムメモリ１１２０または別の種類のメモリとしてのコンピュータシステム１１００のいくつかの実施形態中に備えられていてもよい。さらに、コンピュータアクセス可能な媒体は、ネットワークインターフェース１１４０を介して実行されてもよいような、ネットワーク及び／または無線リンクなどの通信媒体を介して伝達される、電気信号、電磁信号、もしくはデジタル信号などの伝送媒体または信号を挙げることができる。

【０１１２】

本明細書において議論または示唆されるさまざまな実施形態は、多種多様な動作環境で実行することができ、場合によっては、多数のアプリケーションのいずれかを動作させるために使用することができる１つ以上のユーザコンピュータ、計算デバイス、または処理デバイスを含んでいてもよい。ユーザまたは顧客デバイスは、標準的なオペレーティングシステムを実行するデスクトップコンピュータまたはラップトップコンピュータなどの多数の汎用パーソナルコンピュータ、ならびにモバイルソフトウェアを実行し、多数のネットワーキング及びメッセージングプロトコルをサポートすることができるセルラー、ワイヤレス、及びハンドヘルドデバイスのいずれかを備えていてもよい。かかるシステムは、開発及びデータベース管理などの目的で、さまざまな市販のオペレーティングシステムや他の既知のアプリケーションを実行する多数のワークステーションも備えていてよい。これらのデバイスは、ダミー端末、シンクライアント、ゲームシステム、及び／またはネットワークを介して通信する能力を有する他のデバイスなどの他の電子デバイスも備えていてよい。

【０１１３】

ほとんどの実施形態は、Ｔｒａｎｓｍｉｓｓｉｏｎ Ｃｏｎｔｒｏｌ Ｐｒｏｔｏｃｏｌ／Ｉｎｔｅｒｎｅｔ Ｐｒｏｔｏｃｏｌ （ＴＣＰ／ＩＰ）、Ｆｉｌｅ Ｔｒａｎｓｆｅｒ Ｐｒｏｔｏｃｏｌ （ＦＴＰ）、Ｕｎｉｖｅｒｓａｌ Ｐｌｕｇ ａｎｄ Ｐｌａｙ （ＵＰｎＰ）、Ｎｅｔｗｏｒｋ Ｆｉｌｅ Ｓｙｓｔｅｍ （ＮＦＳ）、Ｃｏｍｍｏｎ Ｉｎｔｅｒｎｅｔ Ｆｉｌｅ Ｓｙｓｔｅｍ （ＣＩＦＳ）、Ｅｘｔｅｎｓｉｂｌｅ Ｍｅｓｓａｇｉｎｇ ａｎｄ Ｐｒｅｓｅｎｃｅ Ｐｒｏｔｏｃｏｌ （ＸＭＰＰ）、ＡｐｐｌｅＴａｌｋなどの、広く利用可能なさまざまなプロトコルのいずれかを使用する通信をサポートするために、当業者によく知られている少なくとも１つのネットワークを利用する。上記ネットワーク（複数可）としては、例えば、ローカルエリアネットワーク（ＬＡＮ）、ワイドエリアネットワーク（ＷＡＮ）、仮想プライベートネットワーク（ＶＰＮ）、インターネット、イントラネット、エクストラネット、公衆交換電話網（ＰＳＴＮ

）、赤外線ネットワーク、無線ネットワーク、及びそれらの組み合わせなどを挙げることができる。

【０１１４】

ウェブサーバーを利用する実施形態において、上記ウェブサーバーは、ＨＴＴＰサーバー、Ｆｉｌｅ Ｔｒａｎｓｆｅｒ Ｐｒｏｔｏｃｏｌ （ＦＴＰ）サーバー、Ｃｏｍｍｏｎ Ｇａｔｅｗａｙ Ｉｎｔｅｒｆａｃｅ （ＣＧＩ）サーバー、データサーバー、Ｊａｖａサーバー、ビジネスアプリケーションサーバーなどを含むさまざまなサーバーまたはＭｉｄ－Ｔｉｅｒアプリケーションのいずれかを実行することができる。上記サーバー（複数可）は、Ｊａｖａ（登録商標）、Ｃ、Ｃ＃もしくはＣ＋＋、またはＰｅｒｌ、Ｐｙｔｈｏｎ、ＰＨＰ、もしくはＴＣＬなどのいずれかのスクリプト言語、ならびにそれらの組み合わせなどの任意のプログラミング言語で書かれた、１つ以上のスクリプトあるいはプログラムとして実装することができる１つ以上のＷｅｂアプリケーションを実行することなどにより、ユーザーデバイスからの要求に応答してプログラムまたはスクリプトを実行する能力も有していてよい。上記サーバー（複数可）としては、Ｏｒａｃｌｅ（登録商標）、Ｍｉｃｒｏｓｏｆｔ（登録商標）、Ｓｙｂａｓｅ（登録商標）、ＩＢＭ（登録商標）などから市販されているものを含む、但しこれらに限定されないデータベースサーバーも挙げることができる。上記データベースサーバーは、リレーショナルまたは非リレーショナル（「ＮｏＳＱＬ」など）データベースサーバー、分散型または非分散型データベースサーバーなどであってよい。

【０１１５】

本明細書に開示される環境は、上述のように、さまざまなデータストアならびに他のメモリ及び記憶媒体を備えていてもよい。これらは、１つ以上のコンピュータに対してローカルな（及び／もしくは１つ以上のコンピュータ中に内在する）、またはネットワーク全体のいずれかもしくはすべてのコンピュータから遠隔のストレージメディア上などの、さまざまな場所に存在していてもよい。特定の一連の実施形態において、情報は、当業者によく知られたストレージエリアネットワーク（ＳＡＮ）中に常駐していてもよい。同様に、コンピュータ、サーバー、または他のネットワークデバイスに起因する機能を実行するために必要ないずれかのファイルは、ローカル及び／またはリモートに適切に格納されてもよい。システムがコンピュータ化されたデバイスを備える場合には、それぞれのかかるデバイスは、バスを介して電気的に連結している場合があるハードウェアの構成要素を備えていてもよく、該構成要素としては、例えば、少なくとも１つの中央演算装置（ＣＰＵ）、少なくとも１つの入力装置（例えば、マウス、キーボード、コントローラ、タッチスクリーン、もしくはキーパッド）、及び／または少なくとも１つの出力装置（例えば、ディスプレイ装置、プリンタ、もしくはスピーカー）が挙げられる。かかるシステムは、ディスクドライブ、光学式記憶装置、及びランダムアクセスメモリ（ＲＡＭ）または読み取り専用メモリ（ＲＯＭ）などのソリッドステート記憶装置、ならびに脱着可能なメディアデバイス、メモリカード、フラッシュカードなどの１つ以上の記憶装置も備えていてよい。

【０１１６】

かかる装置は、コンピュータ可読記憶媒体リーダー、通信装置（例えば、モデム、ネットワークカード（無線または有線）、赤外線通信装置など）、及び上述の作業メモリも備えていてよい。上記コンピュータ可読記憶媒体リーダーは、リモート、ローカル、固定、及び／もしくは脱着可能な記憶装置、ならびに一時的及び／もしくはより永続的に、コンピュータ可読情報を収納する、格納する、送信する、及び取得するための記憶媒体を表すコンピュータ可読記憶媒体と接続されていてもよく、またはコンピュータ可読記憶媒体を受信するように構成されていてもよい。上記システム及びさまざまなデバイスは、通常、オペレーティングシステム及び、クライアントアプリケーションまたはウェブブラウザなどのアプリケーションプログラムを含む、少なくとも１つの作業記憶デバイス内に位置する多数のソフトウェアアプリケーション、モジュール、サービス、または他の構成要素も備えることになる。これに代わる実施形態は、上述の実施形態からの多くの変化形を有し

ていてもよいことを理解されたい。例えば、カスタマイズされたハードウェアが使用される場合、及び／またはハードウェア、ソフトウェア（アプレットなどのポータブルソフトウェアを含む）、またはその両方に特定の構成要素が実装される場合もある。さらに、ネットワーク入／出力装置などの他の計算装置への接続が使用される場合もある。

【０１１７】

コードまたはコードの一部を収納するための記憶媒体及びコンピュータ可読媒体としては、コンピュータ読み取り可能な命令、データ構造、プログラムモジュール、もしくはその他のデータなどの情報の格納及び／または送信のための任意の方法または技術に実装される、ＲＡＭ、ＲＯＭ、電気的消去再書込み可能な読出し専用メモリ（ＥＥＰＲＯＭ）、フラッシュメモリまたはその他のメモリ技術、コンパクトディスク読み取り専用メモリ（ＣＤ－ＲＯＭ）、デジタルバーサタイルディスク（ＤＶＤ）もしくはその他の光ストレージ、磁気カセット、磁気テープ、磁気ディスクストレージもしくはその他の磁気記憶装置、または所望の情報を格納するために使用することができ、且つシステムデバイスによってアクセスすることができる任意のその他の媒体を含む、揮発性及び不揮発性、着脱可能な及び着脱不能な記憶媒体などの、但しこれらに限定されない、記憶媒体及び通信媒体を含む、当技術分野で公知のあるいは使用されている任意且つ適宜の媒体を挙げることができる。当業者であれば、本明細書で提供される開示及び教示に基づいて、さまざまな実施形態を実施するための他の手段及び／または方法を理解しよう。

【０１１８】

前述の説明ではさまざまな実施形態が記載されている。説明の目的で、実施形態を完全に理解できるように、特定の構成及び詳細が記載される。しかしながら、当業者には、特定の詳細がなくとも、上記実施形態を実施することができることも明らかであろう。さらに、記載される実施形態を分かりにくくしないために、周知の特徴が割愛または簡略化される場合がある。

【０１１９】

本明細書では、括弧で囲まれた本文及び破線の境界線（例えば、長い破線、短い破線、一点鎖線、及び点線）を伴うブロックは、いくつかの実施形態にさらなる特徴を追加する任意選択の操作を示すために使用される。但し、かかる表記は、これらが唯一の選択肢または任意選択の操作であること、及び／または特定の実施形態において、実線の境界線を伴うブロックが任意選択ではないことを意味すると解釈されるべきものではない。

【０１２０】

Ｖ．免疫原性組成物

本発明はさらに、本明細書に記載の方法を使用して選択される１種以上の腫瘍特異的抗原を含む、個別化された（すなわち、対象特異的な）免疫原性組成物（例えば、がんワクチン）に関する。かかる免疫原性組成物は、当技術分野の標準的な手順に従って処方することができる。本免疫原性組成物は特異的な免疫応答を引き起こす能力を有する。

【０１２１】

本免疫原性組成物は、腫瘍特異的ネオアンチゲンの選択及び数が対象の特定のがんに個別対応するように処方することができる。例えば、上記腫瘍特異的ネオアンチゲンの選択は、特定のがんの種類、がんのステージ、対象の免疫状態、及び対象のＭＨＣ型に依存したものとすることができる。

【０１２２】

本免疫原性組成物は、少なくとも１種、２種、３種、４種、５種、６種、７種、８種、９種、１０種、１１種、１２種、１３種、１４種、１５種、１６種、１７種、１８種、１９種、２０種、２１種、２２種、２３種、２４種、２５種、２６種、２７種、２８種、２９種、３０種、３１種、３２種、３３種、３４種、３５種、３６種、３７種、３８種、３９種、４０種、４１種、４２種、４３種、４４種、４５種、４６種、４７種、４８種、４９種、５０種、またはそれ以上の腫瘍特異的ネオアンチゲンを含んでいてもよい。本免疫原性組成物は、約１０～２０種の腫瘍特異的ネオアンチゲン、約１０～３０種の腫瘍特異的ネオアンチゲン、約１０～４０種の腫瘍特異的ネオアンチゲン、約１０～５０種の腫瘍

特異的ネオアンチゲン、約１０～６０種の腫瘍特異的ネオアンチゲン、約１０～７０種の腫瘍特異的ネオアンチゲン、約１０～８０種の腫瘍特異的ネオアンチゲン、約１０～９０種の腫瘍特異的ネオアンチゲン、または約１０～１００種の腫瘍特異的ネオアンチゲンを含んでいてもよい。本免疫原性組成物は、少なくとも約１０種の腫瘍特異的ネオアンチゲン、または少なくとも約２０種の腫瘍特異的ネオアンチゲンを含むことが好ましい。

【０１２３】

本免疫原性組成物は、天然の抗原または合成された抗原をさらに含むことができる。上記天然または合成の抗原は免疫応答を高める可能性がある。例示的な天然または合成の抗原としては、汎ＤＲエピトープ（ＰＡＤＲＥ）及び破傷風毒素抗原が挙げられるが、これらに限定はされない。

【０１２４】

本免疫原性組成物は、任意の形態、例えば、合成の長鎖ペプチド、ＲＮＡ、ＤＮＡ、細胞、樹状細胞、ヌクレオチド配列、ポリペプチド配列、プラスミド、またはベクターであってもよい。

【０１２５】

腫瘍特異的ネオアンチゲンは、ワクシニア、鶏痘、自己複製アルファウイルス、マラバウイルス、アデノウイルス（Ｔａｔｓｉｓ ｅｔ ａｌ．，Ｍｏｌｅｃｕｌａｒ Ｔｈｅｒａｐｙ，１０：６１６－６２９ （２００４））を参照のこと）、または第２世代、第３世代、もしくはハイブリッド第２／第３世代レンチウイルス、及び特定の細胞型もしくは受容体を標的とするように設計された任意の世代の組換えレンチウイルスを含む、但しこれらに限定されないレンチウイルス（例えば、Ｈｕ ｅｔ ａｌ．，Ｉｍｍｕｎｏｌ Ｒｅｖ．，２３９（１）：４５－６１ （２０１１）；Ｓａｋｍａ ｅｔ ａｌ，Ｂｉｏｃｈｅｍ Ｊ．，４４３（３）：６０３－１８ （２０１２）を参照のこと）などの、ウイルスベクターベースのワクチンプラットフォームに含めることもできる。上述のウイルスベクターベースのワクチンプラットフォームのパッケージング容量に応じて、この手法は、１種以上の腫瘍特異的ネオアンチゲンペプチドをコードする１種以上のヌクレオチド配列を送達することができる。上記配列には非変異配列が隣接していてもよく、リンカーによって分離されていてもよく、または細胞内区画を標的とする１つ以上の配列が先行していてもよい（例えば、Ｇｒｏｓ ｅｔ ａｌ．，Ｎａｔ Ｍｅｄ．，２２（４）：４３３－８ （２０１６）；Ｓｔｒｏｎｅｎ ｅｔ ａｌ．，Ｓｃｉｅｎｃｅ．，３５２（６２９１）：１３３７－１３４１ （２０１６）；Ｌｕ ｅｔ ａｌ．，Ｃｌｉｎ Ｃａｎｃｅｒ Ｒｅｓ．，２０（１３）：３４０１－３４１０ （２０１４）を参照のこと）。宿主に導入されると、感染した細胞は当該の１種以上の腫瘍特異的ネオアンチゲンを発現し、それによって該１種以上の腫瘍特異的ネオアンチゲンに対する宿主免疫（例えば、ＣＤ８＋またはＣＤ４＋）応答を誘発する。免疫化プロトコルに有用なワクシニアベクター及び方法は、例えば、米国特許第４，７２２，８４８号に記載されている。別のベクターとしてはＢＣＧ（Ｂａｃｉｌｌｅ Ｃａｌｍｅｔｔｅ Ｇｕｅｒｉｎ）がある。ＢＣＧベクターはＳｔｏｖｅｒ ｅｔ ａｌ．（Ｎａｔｕｒｅ ３５１：４５６－４６０ （１９９１））に記載される。本明細書の記載から当業者には明らかであろう、ネオアンチゲンの治療上の投与または免疫化に有用な多種多様の他のワクチンベクターも使用することができる。

【０１２６】

本免疫原性組成物は、特定の対象の個人的なニーズに応じて、個別化された成分を含むことができる。

【０１２７】

本明細書に記載の免疫原性組成物は、アジュバントをさらに含んでいてもよい。アジュバントとは、それを免疫原性組成物中に混合することにより、腫瘍特異的ネオアンチゲンに対する免疫応答が増加する、または他の形態で増強及び／またはブーストされる任意の物質であるが、当該物質が単独で投与される場合には、腫瘍特異的ネオアンチゲンに対する免疫応答は生じない。上記アジュバントは、当該ネオアンチゲンに対する免疫応答は生

じさせるが、アレルギーやその他の有害反応を起こさないことが好ましい。本明細書では、本免疫原性組成物は、当該免疫原性組成物の投与の前に、投与と共に、投与と同時に、または投与の後に投与されてもよいことが企図される。

【０１２８】

アジュバントは、例えば、リンパ球のリクルート、Ｂ細胞及び／またはＴ細胞の刺激、ならびにマクロファージの刺激を含む、いくつかの機序によって免疫応答を増強することができる。本発明の免疫原性組成物がアジュバントを含むか、または１種以上のアジュバントと共に投与される場合、使用することができるアジュバントとしては、ミネラル塩アジュバントまたはミネラル塩ゲルアジュバント、粒子アジュバント、微粒子アジュバント、粘膜アジュバント、及び免疫刺激アジュバントが挙げられるが、これらに限定はされない。アジュバントの例としては、アルミニウム塩（ミョウバン）（水酸化アルミニウム、リン酸アルミニウム、及び硫酸アルミニウムなど）、３－Ｏ－脱アシル化モノホスホリル脂質Ａ（ＭＰＬ）（ＧＢ２２２０２１１を参照のこと）、ＭＦ５９（Ｎｏｖａｒｔｉｓ）、ＡＳ０３（Ｇｌａｘｏ　ＳｍｉｔｈＫｌｉｎｅ）、ＡＳ０４（Ｇｌａｘｏ　ＳｍｉｔｈＫｌｉｎｅ）、ポリソルベート８０（Ｔｗｅｅｎ　８０；ＩＣＬ　Ａｍｅｒｉｃａｓ，Ｉｎｃ．）、イミダゾピリジン化合物（国際公開第ＷＯ２００７／１０９８１２号として公開された国際出願第ＰＣＴ／ＵＳ２００７／０６４８５７）、イミダゾキノキサリン化合物（国際公開第ＷＯ２００７／１０９８１３号として公開された国際出願第ＰＣＴ／ＵＳ２００７／０６４８５８号を参照のこと）、及びＱＳ２１などのサポニン（Ｋｅｎｓｉｌ ｅｔ ａｌ，ｉｎ Ｖａｃｃｉｎｅ Ｄｅｓｉｇｎ：Ｔｈｅ Ｓｕｂｕｎｉｔ ａｎｄ Ａｄｊｕｖａｎｔ Ａｐｐｒｏａｃｈ（ｅｄｓ．Ｐｏｗｅｌｌ & Ｎｅｗｍａｎ，Ｐｌｅｎｕｍ Ｐｒｅｓｓ，ＮＹ，１９９５）；米国特許第５，０５７，５４０号を参照のこと）が挙げられるが、これらに限定はされない。いくつかの実施形態において、上記アジュバントはフロイントアジュバント（完全または不完全）である。他の好適なアジュバントは、任意選択でモノホスホリル脂質Ａなどの免疫刺激剤と組み合わせた、水中油型乳化液（スクアレンまたは落花生油など）である（Ｓｔｏｕｔｅ ｅｔ ａｌ．，Ｎ．Ｅｎｇｌ．Ｊ．Ｍｅｄ．３３６，８６－９１（１９９７）を参照のこと）。

【０１２９】

ＣｐＧ免疫賦活性オリゴヌクレオチドが、ワクチン環境においてアジュバントの効果を高めることも報告されている。他のＴＬＲに結合する分子、例えばＴＬＲ７、ＴＬＲ８、及び／またはＴＬＲ９に結合するＲＮＡも使用することができる。

【０１３０】

有用なアジュバントの他の例としては、化学修飾ＣｐＧ（例えば、ＣｐＲ、Ｉｄｅｒａ）、Ｐｏｌｙ（Ｉ：Ｃ）（例えば、ｐｏｌｙｉ：ＣＩ２Ｕ）、Ｐｏｌｙ　ＩＣＬＣ、非ＣｐＧ細菌ＤＮＡまたはＲＮＡ、ならびに、治療的に及び／またはアジュバントとして作用する可能性がある免疫活性小分子及び抗体、例えば、シクロホスファミド、スニチニブ、ベバシズマブ、セレブレクス（セレコキシブ）、ＮＣＸ－４０１６、シルデナフィル、タダラフィル、バルデナフィル、ソラフィニブ、ＸＬ－９９９、ＣＰ－５４７６３２、パゾパニブ、ＺＤ２１７１、ＡＺＤ２１７１、イピリムマブ、トレメリムマブ、及びＳＣ５８１７５が挙げられるが、これらに限定はされない。実施形態において、Ｐｏｌｙ　ＩＣＬＣが好ましいアジュバントである。

【０１３１】

本免疫原性組成物は、本明細書に記載される１種以上の腫瘍特異的ネオアンチゲンを、単独でまたは薬学的に許容される担体と共に含んでいてもよい。１種以上の腫瘍特異的ネオアンチゲンの懸濁液または分散液、特に等張水性懸濁液、分散液、または両親媒性溶媒を使用することができる。本免疫原性組成物は、滅菌されていてもよく、且つ／または賦形剤、例えば防腐剤、安定剤、湿潤剤及び／もしくは乳化剤、可溶化剤、浸透圧を調節するための塩及び／もしくは緩衝剤を含んでいてもよく、それ自体は公知である方法で、例えば従来の分散及び懸濁プロセスによって調製される。特定の実施形態において、かかる

分散液または懸濁液は粘度調整剤を含んでいてもよい。上記懸濁液または分散液は約２℃～８℃の温度に維持されるか、または、より長期間保管する場合には、凍結し、使用直前に解凍することが優先される。注射の場合には、ワクチンまたは免疫原性製剤は、水溶液、好ましくはハンクス液、リンゲル液、または生理食塩水緩衝液などの生理学的に適合する緩衝液中で製剤化されてもよい。上記の液は、懸濁剤、安定剤、及び／または分散剤などの配合剤を含んでいてもよい。

【０１３２】

特定の実施形態において、本明細書に記載の組成物は、防腐剤、例えば水銀誘導体チメロサールをさらに含む。特定の実施形態において、本明細書に記載の医薬組成物は０．００１％～０．０１％のチメロサールを含む。他の実施形態において、本明細書に記載の医薬組成物は防腐剤を含まない。

【０１３３】

賦形剤はアジュバントとは独立して存在していてもよい。賦形剤の機能は、例えば、本免疫原性組成物の分子量を増加させること、活性もしくは免疫原性を増加させること、安定性を付与すること、生物活性を増加させること、または血清半減期を増加させることであってもよい。賦形剤は、Ｔ細胞（例えば、ＣＤ４＋またはＣＤ８＋　Ｔ細胞）に対する、上記１種以上の腫瘍特異的ネオアンチゲンの提示を補助するために使用されてもよい。上記賦形剤は、キーホールリンペットヘモシアニン；トランスフェリン、ウシ血清アルブミン、ヒト血清アルブミン、サイログロブリンもしくはオボアルブミン、免疫グロブリンなどの血清タンパク質；またはインスリンもしくはパルミチン酸などのホルモンなどの、但しこれらに限定されないキャリアタンパク質であってもよい。ヒトの免疫化の場合、上記担体は一般に、人に対して許容され且つ安全な、生理学的に許容される担体である。あるいは、上記担体は、デキストラン、例えばセファロースであってもよい。

【０１３４】

細胞傷害性Ｔ細胞は、インタクトな外来抗原そのものではなく、ＭＨＣ分子に結合したペプチドの形態で抗原を認識する。上記ＭＨＣ分子は、抗原提示細胞の細胞表面に存在する。したがって、ペプチド抗原、ＭＨＣ分子、及び抗原提示細胞（ＡＰＣ）の三量体複合体が存在する場合、細胞傷害性Ｔ細胞の活性化が可能である。細胞傷害性Ｔ細胞の活性化に１種以上の腫瘍特異的抗原を使用されるだけでなく、それぞれのＭＨＣ分子を含むさらなるＡＰＣが追加されるならば、免疫応答が強化される可能性がある。したがって、いくつかの実施形態において、免疫原性組成物は、少なくとも１種のＡＰＣをさらに含む。

【０１３５】

本免疫原性組成物は、許容可能な担体（例えば、水性担体）を含んでいてもよい。さまざまな水性担体、例えば、水、緩衝水、０．９％生理食塩水、０．３％グリシン、ヒアルロン酸などを使用することができる。これらの組成物は、従来の周知の滅菌技法によって滅菌されてもよく、または滅菌ろ過されてもよい。得られる水溶液は、そのまま使用するように包装されてもよく、または凍結乾燥され、投与前に滅菌溶液と混合されてもよい。上記組成物は、ｐＨ調整剤及び緩衝剤、張度調整剤、湿潤剤などの、生理学的条件に近づけるのに必要な薬学的に許容される補助物質、例えば、酢酸ナトリウム、乳酸ナトリウム、塩化ナトリウム、塩化カリウム、塩化カルシウム、モノラウリン酸ソルビタン、オレイン酸トリエタノールアミンなどを含んでいてもよい。

【０１３６】

ネオアンチゲンはリポソームを介して投与されてもよく、該リポソームによって上記ネオアンチゲンはリンパ組織などの特定の細胞組織を標的化する。リポソームは半減期を増加させるのにも有用である。リポソームとしては、乳化液、泡状物質、ミセル、不溶性単分子膜、液晶、リン脂質分散液、ラメラ層などが挙げられる。送達されるネオアンチゲンは、これらの調製物中に、リポソームの一部として、単独で、あるいは例えば、ＣＤ４５抗原に結合するモノクローナル抗体などの、リンパ系細胞中に広く存在する受容体に結合する分子、または他の治療用組成物もしくは免疫原性組成物と共に組み込まれる。このようにして、所望のネオアンチゲンを充填したリポソームはリンパ系細胞の部位に向かうこ

とができ、次いでリポソームは該部位に選択された免疫原性組成物を送達する。リポソームは標準的なベシクル形成性脂質から形成することができ、該脂質としては一般に、中性及び負に荷電したリン脂質及びコレステロールなどのステロールが挙げられる。脂質の選択は、概括的には、例えば、リポソームのサイズ、酸不安定性、及び血流中での当該リポソームの安定性への考慮を指針として行われる。リポソームの調製には、例えば、Ｓｚｏｋａ ｅｔ ａｌ．, Ａｎ．Ｒｅｖ．Ｂｉｏｐｈｙｓ．Ｂｉｏｅｎｇ．９；４６７（１９８０）、米国特許第４，２３５，８７１号、第４，５０１，７２８号、第４，５０１，７２８号、第４，８３７，０２８号、及び第５，０１９，３６９号に記載されるような、さまざまな方法が利用可能である。

【０１３７】

免疫細胞を標的とするために、リポソーム中に組み込まれるリガンドは、例えば、所望の免疫系細胞の細胞表面決定基に特異的な抗体または抗体のフラグメントを含んでいてもよい。リポソーム懸濁液は、とりわけ、投与方法、送達されるペプチド、及び治療を受ける疾患のステージに応じて変化する用量で、静脈内投与、局所投与（ｌｏｃａｌｌｙ）、局所投与（ｔｏｐｉｃａｌｌｙ）などによって投与されてもよい。

【０１３８】

免疫細胞；抗原（すなわち、腫瘍特異的ネオアンチゲン）、リガンド、またはアジュバント（例えば、ＴＬＲ）などの本免疫原性組成物の成分を標的とするための代替方法は、乳酸－グリコール酸共重合体ミクロスフェア中に組み込まれてもよい。上記乳酸－グリコール酸共重合体ミクロスフェアは、エンドソーム送達デバイスとして本免疫原性組成物の成分を封入することができる。

【０１３９】

治療または免疫化の目的で、本明細書に記載の腫瘍特異的ネオアンチゲンをコードする核酸が対象に投与されてもよい。上記核酸を対象に送達するには、多くの方法が利便性よく使用される。例えば、上記核酸は「裸のＤＮＡ」として直接送達されてもよい。この手法は、例えば、Ｗｏｌｆｆ ｅｔ ａｌ．, Ｓｃｉｅｎｃｅ ２４７：１４６５－１４６８（１９９０）、ならびに米国特許第５，５８０，８５９号及び第５，５８９，４６６号に記載される。上記核酸は、例えば、米国特許第５，２０４，２５３号に記載される弾道送達を使用して投与されてもよい。ＤＮＡのみからなる粒子が投与されてもよい。あるいは、ＤＮＡは金粒子などの粒子に付着していてもよい。核酸配列を送達するための手法としては、エレクトロポレーションを伴うまたは伴わない、ウイルスベクター、ｍＲＮＡベクター、及びＤＮＡベクターを挙げることができる。上記核酸は、カチオン性脂質などのカチオン性化合物と複合体を形成して送達されてもよい。

【０１４０】

本明細書で提供される免疫原性組成物は、経口、皮内、腫瘍内、筋肉内、腹腔内、静脈内、局所、皮下、経皮、鼻腔内、及び吸入経路を含む、但しこれらに限定されない経路によって、ならびに乱切法（例えば、二股状の針を使用して、皮膚の表層を通してかき傷をつける）によって対象に投与されてもよい。本免疫原性組成物を腫瘍部位に投与して、腫瘍に対する局所免疫応答を誘導してもよい。

【０１４１】

上記１種以上の腫瘍特異的ネオアンチゲンの用量は、組成物の種類、及び対象の年齢、体重、体表面積、個々の状態、個々の薬物動態データ、及び投与方法に依存する場合がある。

【０１４２】

本明細書では、本明細書に開示の方法のステップを実施することによって選択される、１種以上の腫瘍特異的ネオアンチゲンを含む免疫原性組成物の製造方法も開示される。本明細書に記載の免疫原性組成物は、当技術分野で公知の方法を使用して製造することができる。例えば、本明細書に開示の腫瘍特異的ネオアンチゲンまたはベクター（例えば、１種以上の腫瘍特異的ネオアンチゲンをコードする少なくとも１つの配列を含むベクター）の製造方法は、上記ネオアンチゲンもしくはベクターを発現するのに好適な条件下で、上

記ネオアンチゲンまたはベクターをコードする少なくとも１種のポリヌクレオチドを含む宿主細胞を培養することと、上記ネオアンチゲンまたはベクターを精製することを含んでいてもよい。標準的な精製方法としては、クロマトグラフィー技法、電気泳動技法、免疫学的技法、沈殿技法、透析技法、ろ過技法、濃縮技法、及びクロマトフォーカシング技法が挙げられる。

【０１４３】

宿主細胞としては、チャイニーズハムスター卵巣（ＣＨＯ）細胞、ＮＳ０細胞、酵母、またはＨＥＫ２９３細胞を挙げることができる。宿主細胞は、本明細書に開示の１種以上の腫瘍特異的ネオアンチゲンまたはベクターをコードする少なくとも１つの核酸配列を含む１種以上のポリヌクレオチドで形質転換されていてもよい。特定の実施形態において、単離されたポリヌクレオチドはｃＤＮＡであってよい。

【実施例】

【０１４４】

例１．１：トレーニングデータ

モデルを、ペプチド－ＭＨＣ結合及びＭＨＣクラスＩ上での内因性ペプチド提示の確率を予測するようにトレーニングを行った。これらは、ＣＤ８＋Ｔ細胞免疫原性の代用値として扱う。ＭＨＣｆｌｕｒｒｙから精選されたペプチド－ＭＨＣ結合親和性データ（“ｃｕｒａｔｅｄ＿ｔｒａｉｎｉｎｇ＿ｄａｔａ．ｎｏ＿ｍａｓｓ＿ｓｐｅｃ．ｃｓｖ”１）を使用し、これはＩＥＤＢ［１］及びＫｉｍ　ｅｔ　ａｌ．［２］からのデータを含んでいる。この精選されたデータセットに対して実施される唯一の処理ステップは、隣接領域を抽出し、ネガティブサンプリング法に使用するためにペプチドのソースタンパク質（複数可）を追加することであった。この処理ステップの後、トレーニングに使用される最終的なデータセットは“ｃｕｒａｔｅｄ＿ｔｒａｉｎｉｎｇ＿ｄａｔａ．ｎｏ＿ｍａｓｓ＿ｓｐｅｃ．ｍｕｌｔｉｐｌｅ＿ｃｏｎｔｅｘｔ．ｂｌａｓｔ．ｖ２．ｃｓｖ．”と称される。このデータセットからは、ＨＬＡ－Ａ／Ｂ／Ｃ対立遺伝子を含むエントリ、及び翻訳後修飾のない長さ８～１５のペプチドのみが保持された。これらのサンプルの目標は、ＭＨＣｆｌｕｒｒｙ－１．２［３］で提案されているように、定量的（「＝」）または定性的（「＜」／「＞」）のいずれかである。ＭＨＣｆｌｕｒｒｙの精選されたデータセットの定性的エントリは、陽性－高（＜１００ｎｍ）、陽性－中（＜１０００ｎｍ）、陽性－低（＜５０００ｎｍ）、または陰性（＞５０００ｎｍ）という定性的な値を表す。さらに、いくつかのソース由来の、ペプチド溶離実験及び質量分析によって測定された細胞表面に提示されたペプチドからなる細胞表面へのペプチド提示データセット（Ｓａｒｋｉｚｏｖａ　ｅｔ　ａｌ．［４］データセット、質量分析法を使用して、９５種のＨＬＡ－Ａ型、Ｂ型、またはＣ型、及びＧ型の単一対立遺伝子細胞株から溶離した、１８５，０００種を超えるペプチドをプロファイリングを実施）を利用した。ＭＨＣｆｌｕｒｒｙの精選されたデータセットにおいて、質量分析によって特定された場合に質量分析が的中しており、該データセットにおいて、関連するサンプルは、「ｍｅａｓｕｒｅｍｅｎｔ＿ｓｏｕｒｃｅ」列中に「質量分析」値が記載されていることにより特定される。これは、ＩＥＤＢ［１］もしくはＳｙｓｔｅＭＨＣ　Ａｔｌａｓ［５］に寄託された、またはＡｂｅｌｉｎ　ｅｔ　ａｌ．［６］によって報告された２２６，６８４種のＭＳによって特定されたリガンドを含んでいる。

【０１４５】

さらに、Ｆｒｅｄ　Ｈｕｔｃｈｉｎｓｏｎ　Ｃａｎｃｅｒ　Ｃｅｎｔｅｒにおいて、ＨＥＫ２９３細胞を安定的にトランスフェクトして、β－２－ミクログロブリンに共有結合した分泌型ＨＬＡ分子を発現させることによって取得した細胞表面提示データを使用した。次いで、細胞上清を調製して、分泌されたＭＨＣペプチド複合体を捕捉し、質量分析法で分析した。精製した複合体に由来するペプチドについて報告がなされていた。そのデータファイルは、圧縮ファイル“ＲｏｌａｎｄＰｅｐｔｉｄｅＰｒｅｓｅｎｔａｔｉｏｎＤａｔａ．ｚｉｐ”で入手可能である。これらのデータソースはすべて、特定のＨＬＡ対立遺伝子と関連するＭＨＣクラスＩ分子上で提示されることが判明しているペプチド配列で

構成される。本発明者らの最終的な提示データセット（“ｍａｓｓ＿ｓｐｅｃ＿ｄａｔａ．ｍｕｌｔｉｐｌｅ＿ｃｏｎｔｅｘｔ．ｂｌａｓｔ．ａｌｌｅｌｅ＿ｓｕｐｅｒｔｙｐｅｓ．ｖ３．ｃｓｖ”）に関しては、ＨＬＡ－Ａ型／Ｂ型／Ｃ型対立遺伝子及びペプチドのソースタンパク質（複数可）を含むすべての言及したソースのサンプルを、隣接領域を抽出し、ネガティブサンプリング法に使用するために合体させた。同一の対立遺伝子及び同一のペプチドを有し、いずれかの縁部に単一の追加のアミノ酸を有するのみのサンプルは、質量分析測定の不正確さに起因して重複と見なされることから、重複に近い「拡張」サンプルもフィルタリングにより除外した。上記免疫原性評価データ中に見られるサンプル（ペプチド－ＭＨＣ対）も、親和性データセット及び提示データセットの両方からフィルタリングにより除外し、トレーニング段階で、上記評価セットは完全に「隠され」、モデルに対して全く示されないことを確認した。

【０１４６】

これらのデータセット（親和性及び提示）を、新しいモデルをトレーニングする前に、トレーニング分割部分と検証分割部分にランダムに分割した。各対立遺伝子について、ｉ番目の対立遺伝子の結合親和性サンプルからN_{bai}ペプチドをランダムにサンプリングし、ｉ番目の対立遺伝子の提示サンプルからN_{pi}ペプチドをランダムにサンプリングする。ここで、N_{bai}＝ｍｉｎ（０．２５＊｜固有の親和性ペプチド｜ｉ，１００）及びN_{pi}＝ｍｉｎ（０．２５＊｜固有の提示ペプチド｜$_i$，１００）である。

【０１４７】

各対立遺伝子について、これらのサンプリングされたペプチド（両方のデータセットからサンプリングされたペプチドの合体）は隠された検証セットとみなされ、これらのペプチドを含むすべてのサンプルはトレーニングセットから除去され、検証のみに使用される。上記検証セットは、Ｎ回を超えて連続するエポックにおいて、検証損失が好転しない場合に、モデルトレーニングの早期の停止を決定するために使用される。すべての実験でＮ＝２０を設定した。複数の類似のモデルをトレーニングする場合、アンサンブル／モデル選択目的で、異なるトレーニング－検証分割部分を使用する。

【０１４８】

ヒトライノウイルス（ＨＲ）データを、ＨＲＶ－１Ａ、ＨＲＶ－Ｂ、ＨＲＶ－Ｃモザイクから構築された１６００　ＨＲＶ　１５量体に適用したＩＣＳからの陰性サンプルとして（Ｆｉｓｃｈｅｒ　ｅｔ　ａｌ．，（２００７），Ｎａｔｕｒｅ　Ｍｅｄｉｃｉｎｅ，１３，１００－１０６に記載）、本発明者らのデータに含めた。詳細には、ＩＣＳの結果からランダムにサンプリングした陰性サンプルを免疫原性評価に使用し、免疫原性評価データの一部として試験した。残余のＨＲＶサンプルは陰性提示サンプルとして扱い、トレーニング中にランダムにサンプリングした。

【０１４９】

例１．２：免疫原性評価データ

ＢｉｇＭＨＣ　１．０によるペプチド－ＭＨＣ結合親和性及び細胞表面へのペプチド提示の予測が、Ｔ細胞免疫原性の予測にどの程度十分に変換可能かを検証するために、Ｔ細胞免疫原性データセットを作成し、ＢｉｇＭＨＣの機械学習モデルを試験及び検証した。ＨＩＶ分子免疫学データベースにおけるＣＴＬ／ＣＤ８＋エピトープの要約表、及び、既報の、ＨＣＶ免疫学データベースにおけるＣＴＬエピトープの要約表由来の、既報のペプチド－ＭＨＣ対を使用した。これらの表は、実験的に検証されたＨＩＶ／ＨＣＶ　ＣＴＬ／ＣＤ８＋エピトープを提供する。

【０１５０】

これらのサンプルはすべて免疫原性が陽性のクラス中にある。同様に免疫原性が陰性のペプチド－ＭＨＣサンプルを得るために、陽性の対が生み出される実験を手作業で再検討した。否定的な知見は報告されていない一方、それらの一部は肯定的な知見から再構成することができることが判明した。詳細には、多くの実験において、所定のペプチドのセットと所定のＨＬＡ対立遺伝子のセットについて、すべての可能なペプチド－ＭＨＣ対の組み合わせを試験したが、これはマトリクス法と呼ばれる。肯定的な知見から既報のペプチ

ド及び対立遺伝子を抽出することができ、少なくとも（すべての可能な組み合わせのうちでペプチド／対立遺伝子が陰性であるかのように、本発明者らが見逃しており、且つ報告されることがない陰性サンプルがさらに存在する可能性がある）、上記ペプチド及び対立遺伝子内の可能性のある全てのペプチド－ＭＨＣの対をなす組み合わせを、上記実験において試験したと結論付けることができた。この試験したペプチド－ＭＨＣ対の一覧を考慮すると、この一覧にある陽性として報告されなかった対はいずれも、実際に陰性であると結論付けることができる。マトリクス法が使用されたかどうかを検証するために、３１の最大規模の実験（最大量の試験サンプルを使用、｜固有の対立遺伝子｜ｘ｜固有のペプチド｜）を再検討したところ、そのうちの１８でこの方法が使用されていた。他の方法を使用した実験からは、陰性サンプルは推測されなかった。

【０１５１】

残余のより小規模な実験については、マトリクス法が使用され、サンプルが別の実験で陽性と報告されていない限り、すべての陰性サンプルが抽出されたと想定され、サンプルが別の実験で陽性と報告された場合は、本発明者らは、該サンプルは陽性であると想定した。ＨＬＡ－Ａ型／Ｂ型／Ｃ型対立遺伝子と、長さが８～１５のペプチドのみが保持された。

【０１５２】

本発明者らは、上記のデータに加えて、ＩＥＤＢで報告された免疫原性陽性サンプル（Ｖｉｔａ ｅｔ ａｌ．，（２０１９），Ｎｕｃｌｅｉｃ ａｃｉｄｓ ｒｅｓｅａｒｃｈ，４７，Ｄ３３９－Ｄ３４３）及びランダムにサンプリングされた追加のＨＲＶ陰性サンプルを追加する。陽性サンプル：陰性サンプル比が比較的低い対立遺伝子については、ＨＲＶ陰性サンプルをサンプリングする。本発明者らは理想的な比率は１：１００であると考えており、可能であれば（すべての対立遺伝子についてＨＲＶサンプルを有しているわけではない）、ほぼこの比率に到達するまでＨＲＶ陰性サンプルをサンプリングする。この均衡化手順に従って、１：５よりも小さい比率の対立遺伝子のサンプルをすべてフィルタリングにより除外する。

【０１５３】

この手順に従って、１１０個のＨＬＡ対立遺伝子と１，４１６個の固有のペプチドをカバーする、２，９８５個の陽性サンプルと６８，４６９個の陰性サンプルの対が得られた。

【０１５４】

この免疫原性データセットを２つのセット、すなわち、本発明者らのモデルのハイパーパラメータのチューニング及びさらなる構成の選択を行うための検証セット６と、本発明者らが最終的なベンチマーキングを実行する対象である試験セット７に分割した。

【０１５５】

例１．３：データ解析

上記データの分布をより良く視覚化及び理解するために、次の図をプロットした。

【０１５６】

ペプチドの長さ：図２のＢは３つのデータセットすべてにわたるペプチド長の分布を示す。

【０１５７】

ペプチドの多様性：２つのペプチド間の類似性は、可能な最良のアラインメントで重複するアミノ酸の数と呼ばれる。図２のＡにおいて、各類似性閾値について、所定の閾値を超える類似性を有するパートナーペプチドを持たないペプチドの割合を計算した。この解析では、固有のペプチドのみを考慮し、各データセット内で重複するペプチドを無視した。詳細には、親和性データセットは、全体で１５８，００１個のサンプルのうち３５，４６７個が固有のペプチドで構成され（２２．４５％）、提示データセットは、全体で３８４，８１２個のサンプルのうち２６５，２３６個が固有のペプチドで構成され（６８．９３％）、免疫原性データセットは、全体で７１，４７４個のサンプルのうち１４１６個が固有のペプチドで構成されていた（１．９８％）。

【０１５８】
目標の分布：目標の分布及びそれらの間の不均衡は、異なるデータセット間で異なっていた。すなわち、結合親和性 － ペプチド－ＭＨＣ結合親和性データセットは、定量的目標と定性的目標の混合物で構成されていた。図３は結合親和性目標の高レベルの定性的な分布を示す。提示 － 細胞表面へのペプチド提示データセットは二値であり、本発明者らのデータセットは陽性サンプルのみで構成されていた。本発明者らは、トレーニング中に、各エポックの開始時にネガティブサンプルマイニングを適用して、各陽性サンプル毎に対応する陰性サンプルを生成させた。免疫原性 － Ｔ細胞免疫原データセットは二値であるが、非常に不均衡であり、２，９８５個の陽性ペプチド及び６８，４６９個の陰性ペプチド－ＭＨＣ対で構成されていた。

【０１５９】
スーパータイプ毎のサンプル分布 － 対立遺伝子毎に視覚化するのは難しい、データセット中の基礎となる対立遺伝子分布をより適切に視覚化して理解するために、図４は、本発明者らのデータセットサンプルにおけるＨＬＡ対立遺伝子スーパータイプの分布を示す。Ｓｉｄｎｅｙ ｅｔ ａｌ．［７］によって決定されたＨＬＡスーパータイプ分類を適用した。

【０１６０】
ＨＬＡ対立遺伝子毎のサンプル分布 － ＨＬＡ対立遺伝子スーパータイプの分布に加えて、図４は、データセットのペプチド－ＭＨＣサンプルのそれぞれにおけるＨＬＡ対立遺伝子の分布も示す。

【０１６１】
例１．４：トレーニング中のネガティブマイニングデータ
細胞表面へのペプチド提示データは「陽性」サンプルのみで構成されるが、上記陽性サンプルは、二値提示分類器のトレーニングを行うのに必要な陰性サンプル（細胞表面上に提示されることができない）を提供しない。したがって、かかる分類器のトレーニングを行うために、トレーニング中の確率的ネガティブマイニングのための以下の戦略を採用した。

【０１６２】
ＨＬＡ対立遺伝子シャッフル ペプチド及び対応するＨＬＡ対立遺伝子からなる陽性サンプルを所与として、陽性対立遺伝子のスーパータイプ（複数可）に属さない異なる対立遺伝子をランダムにサンプリングすることにより、上記所与の対立遺伝子を置換した。Ｓｉｄｎｅｙ ｅｔ ａｌ．［７］によって決定されたＨＬＡスーパータイプ分類（該分類は各ＨＬＡ対立遺伝子を１種以上のＨＬＡスーパータイプに割り当てる）を適用した。この分類では、いくつかのＨＬＡ対立遺伝子が未分類のまま残る。この未分類の対立遺伝子を、対応するＨＬＡ－Ａ型／Ｂ型／Ｃ型の群に従って、「未分類－Ａ」、「未分類－Ｂ」、及び「未分類－Ｃ」の３つのさらなるスーパータイプクラスにマッピングし、これらの群を他のスーパータイプクラスと同様に扱った。

【０１６３】
ペプチドシャッフル ペプチド及び対応するＨＬＡ対立遺伝子からなる陽性サンプルを所与として、上記所与のペプチドを、該ペプチドのソースタンパク質からランダムにサンプリングした同一の長さのアミノ酸部分配列で置換した。さらに、ＭＨＣｆｌｕｒｒｙ－１．６［８］の方法に従って、定性的な弱い親和性の目標（＞２０，０００ｎＭ）で、アミノ酸データ分布からサンプリングしたランダムペプチドを含むように、親和性トレーニングデータセットも拡張した。これらのランダムペプチドの長さを、各対立遺伝子について、ペプチド長毎に、同数の非結合データポイントを実施するような方法で測定した。

【０１６４】
ＨＲＶネガティブサンプリング 陰性のＨＲＶデータからランダムにサンプリングする（免疫原性評価に使用するサンプルを除く）。ここでの基本的な仮定は、ほとんどの場合、免疫原性が陰性であることは、細胞表面への提示が陰性であることに起因することになり、したがって、トレーニング中にそれを想定することができることである（但し、提示

が陽性であり、免疫原性が陰性の場合も存在する）。
【０１６５】
例１．５：クロスタスクの目標の推測
合同マルチタスクのトレーニングのために、結合親和性データセット及び提示データセットの両方のすべてのトレーニングサンプルを使用した。但し、各サンプルについて、１つのみの既知の目標であって、該目標が由来する、対応するデータセットの上記１つのみの既知の目標（２つのタスクではなく）は既知であった。この問題を軽減するために（且つマルチタスクトレーニングをより効果的に活用できるようにするために）、細胞表面上に提示されるサンプル（提示が陽性）は高い結合親和性の値も有することになり、結合親和性値が低いサンプルは提示されないことになる（提示が陰性）と仮定することにより、それぞれのタスクからもう一方のタスクの目標を推測した。詳細には、提示が陽性のあらゆるサンプルについては、定性的な高親和性の目標（＜５００ｎＭ）を推測し、結合親和性の測定値が低い（＞５０００ｎＭ）サンプルについては提示が陰性の目標を推測した。残余の「見つからない目標」（推測することができない目標）は、ゼロのサンプル重みを割り当てることによって（目標が見つからないタスクについてのみ）、トレーニング中に単純に無視する（覆い隠す）。
【０１６６】
例１．６：自己蒸留
ＢｉｇＭＨＣ予測子を使用して、さまざまなサンプルの結合親和性及び提示の推定を抽出し、対応する「弱い」ラベルを有するこれらのサンプルをトレーニングデータセットに追加した。この自己蒸留プロセスは、以下に示す２つのシナリオで実施した。
【０１６７】
１．複数の対立遺伝子の質量分析データ　ＭＨＣｆｌｕｒｒｙ－２．０に記載のＭＵＬＴＩ－ＡＬＬＥＬＩＣ　ＯＬＤデータセットを使用した。上記データセットは２０万種を超える陽性サンプルの質量分析ヒットを含む。ＢｉｇＭＨＣ－１．３．１予測を利用して、いずれの対立遺伝子がヒットの要因の対立遺伝子であるかを判定した。まず、本発明者らの提示トレーニングデータから、既知の陽性の提示体（ｐｒｅｓｅｎｔｅｒ）との複数の対立遺伝子のヒットをフィルタリングにより除外した。次に、提示の確率が最も高い対立遺伝子を選択し、提示の確率がある閾値（０．５）よりも高く、且つ結合親和性が一定の閾値（５０００ｎＭ）を下回れば、そのサンプルのみを保持する。
【０１６８】
２．陽性の提示体　本発明者らの提示トレーニングデータ中の、結合親和性が未知であるあらゆる陽性の提示体について（上記のステップを実行した後に）、ＢｉｇＭＨＣ－１．３．１予測に基づいて結合親和性を推定した。予測された結合親和性が５０００ｎＭ未満であるすべてのサンプルを結合親和性トレーニングデータに追加した。
【０１６９】
例２．１：配列の表示
各ＨＬＡ対立遺伝子は、ＭＨＣｆｌｕｒｒｙ－１．４で使用される４９アミノ酸の擬似配列によって表されていた。この疑似配列コード化は、種にわたる多数のＭＨＣクラスＩ対立遺伝子の多重配列アラインメントによって決定された４９の選択された位置のアミノ酸を使用する。ＨＬＡ対立遺伝子の擬似配列の表示は“ａｌｌｅｌｅ＿ｓｅｑｕｅｎｃｅｓ．ｃｓｖ”ファイルで利用可能である。さらに、Ｏ’Ｄｏｎｎｅｌｌ　ｅｔ　ａｌ．［３］のペプチドパディング（ｐａｄｄｉｎｇ）及びコード化方法を利用し、最も重要なＭＨＣとの安定化接触を行う残基のポジショナリティーを保存するように設計された固定長コード化を使用して、アミノ酸長８～１５のペプチドを表示した。これらの「アンカー位置」は、ほとんどの対立遺伝子に対して、当該ペプチドの先頭または末尾に向かって生じる。ペプチドは長さ１５の配列として表され、該配列において、欠落した残基は文字「Ｘ」、事実上２１番目のアミノ酸で埋められる。上記ペプチドの最初と最後の４つの残基は、上記表示中の最初と最後の４つの位置にマッピングされる。中央の７残基は必要に応じて埋められる。８量体ではすべての中間位置がＸとして残る一方、１５量体ではすべての位

置が埋められる。このようにして、アンカー残基を含む可能性が最も高い位置が、上記表示中の同一の位置に一貫してマッピングされる。コード化されたペプチドの縁部でそれぞれ連結される片側あたり５アミノ酸を考慮して、上記ペプチドの隣接領域もコード化された。

【０１７０】

ＢＬＯＳＵＭ６２置換行列に基づいて固定化されたアミノ酸の埋め込みを使用するＯ’Ｄｏｎｎｅｌｌ　ｅｔ　ａｌ．［３］とは対照的に、本発明者らのニューラルネットワークの残余と合同して、エンドツーエンド方式でトレーニングが行われる、トレーニング可能な埋め込み層を使用した。この埋め込み層は、上記コード化されたペプチド中または対立遺伝子擬似配列中の両方のあらゆるアミノ酸を１６次元ベクトルにコード化する。

【０１７１】

例２．２：隣接領域

各ペプチド配列について、この配列がＵｎｉＰｒｏｔデータセット［９］中に存在するより長いタンパク質配列のサブセットであるすべてのインスタンスを特定した。３つのＵｎｉＰｒｏｔファイル、すなわち、（１）ＵｎｉＰｒｏｔヒトプロテオームデータセット、“ＵＰ０００００５６４０＿９６０６．ｆａｓｔａ”、（２）完全なＵｎｉＰｒｏｔＫＢ／Ｓｗｉｓｓ－Ｐｒｏｔデータセット、“ｕｎｉｐｒｏｔ＿ｓｐｒｏｔ．ｆａｓｔａ”、（３）すべての注釈付きスプライスバリアントを表す、ＵｎｉＰｒｏｔＫＢ／Ｓｗｉｓｓ－Ｐｒｏｔデータセットの追加配列、“ｕｎｉｐｒｏｔ＿ｓｐｒｏｔ＿ｖａｒｓｐｌｉｃ．ｆａｓｔａ”を検索し、且つ（４）ｎｅｔＭＨＣｐａｎ－４．０免疫原性データセットからの追加配列、ＣＤ８（“ＣＤ８＿ｅｐｉｔｏｐｅｓ＿ｎｅｔＭＨＣｐａｎ．ｆａｓ”）及びＣＤ４（“ＣＤ４＿ｅｐｉｔｏｐｅｓ＿ｎｅｔＭＨＣＩＩｐａｎ．ｆｓａ”）の両方をダウンロードした。

【０１７２】

上記より長い配列のそれぞれは「親配列」と呼ばれる。各ペプチドは、長さ１０の１つ以上の「隣接領域」と関連しており、該隣接領域は、当該ペプチドの親配列のそれぞれにおける、該ペプチドの直前の５アミノ酸、及び該ペプチド配列の直後の５アミノ酸である。隣接領域のすべての固有の組み合わせがファイルに保存され、それぞれのペプチドの固有の配列の数に反比例する、それらのそれぞれの重みを規定した。これらの重みを、可能性のあるすべてのバリエーションをネットワークに学習させながら、多数のバリエーションを有するペプチドをより重視しないようにするための、トレーニング中のサンプルの重みとして使用した。厳密に一致しないペプチドについては、ＢＬＡＳＴ１０を利用して、最もよく一致したペプチドを見つけ、その対応する「親配列」を使用して、関連する隣接領域を抽出した。

【０１７３】

例２．３：自己監視型事前トレーニング

大規模なタンパク質データベースを利用して、本発明者らのモデルを事前トレーニングし、適切な初期配列表示を学習させた。Ｕｎｉｐａｒｃデータベース由来の２５００万種のタンパク質のサブセットを使用して、ＢＥＲＴ事前トレーニングにヒントを得て、以下の２つのタスクでペプチドトランスフォーマーモデルのトレーニングを行った。

【０１７４】

１．マスクされた言語モデリング　「覆い隠された」と見なされるトークンのうち０．１５個をランダムに選択し、交差エントロピー損失を使用して元のトークンを予測しようと試みる（他のすべてのトークンに基づいて）トークン分類ヘッドをトレーニングした。モデル入力については、「覆い隠された」トークンは、ランダムに置換される場合（１０％）、変更されないままになる場合（１０％）、及びマスキングトークンに置換される場合（８０％）がある。

【０１７５】

２．次のペプチドの予測　事前トレーニング段階では、本発明者らの入力配列は２つのペプチド配列の連結であった（主トレーニング段階でのペプチド配列と対立遺伝子配列の

連結とは対照的に）。シーケンスは特別な分離トークン（＜ＳＥＰ＞）を介して分離され、異なるセグメントインデックス及び埋め込みを有する（セグメントシーケンスはネットワークへの追加入力であり、各トークンが第１のシーケンス、第２のシーケンス、または特別なトークンに属するかどうかを示すだけである。次いで、セグメントインデックスの埋め込みがトークンに追加され、埋め込みが配置される）。＜ＣＬＳ＞トークンの出力に対して分類器をトレーニングし、第２のペプチドがタンパク質中で次に生成するペプチド（第１のペプチドの後）であるかどうかを予測した。上記ペプチドは、ヒトタンパク質に由来する２つの連続する、同一の長さのペプチドに由来するか、または異なるタンパク質からランダムにサンプリングされる。

【０１７６】

例２．３　トレーニングオブジェクティブ及びマルチタスク損失

本発明者らのニューラルネットワークに対し、ペプチド－ＭＨＣ結合親和性と細胞表面へのペプチド提示の両方を予測するために合同でトレーニングを行った。Ｏ’Ｄｏｎｎｅｌｌ　ｅｔ　ａｌ．［３］によって実施されたように、ＬＢＡ－ＭＳＥによって示される平均二乗誤差（ＭＳＥ）損失関数の変化形を利用し、これにより、不等式（＞）または（＜）を伴う測定値は、データセット中の定量的及び定性的なペプチド－ＭＨＣ結合親和性測定値の両方の処理に関して、該不等式に反する場合にのみ損失に寄与する。正確な式はスキーム１及びスキーム２に概説されている。

【０１７７】

細胞表面へのペプチド提示の二値分類タスクでは、ＬＰ－ＦＬによって示されるＦｏｃａｌ　Ｌｏｓｓ［１０］を利用し、上記Ｆｏｃａｌ　Ｌｏｓｓは、十分に分類されていないサンプルをより強調する、標準的な二値交差エントロピー損失の加重拡張である。厳密な式をスキーム３に概説している（以下に再掲する）。

【数９】

$$L_{\text{P-FL}} = -\frac{1}{N}\sum_{i=1}^{N}(1-p_{ti})^{\gamma}\log p_{ti}$$

但し、正しいクラスp_{ti}は以下のとおりである。

$$p_{ti} = \begin{cases} \hat{Y}_i & , Y_i = 1 \\ 1-\hat{Y}_i & , Y_I = 0 \end{cases}$$

スキーム３

【０１７８】

ここで、γは実数パラメータであり、１に設定する。二値の場合、

【数１０】

$$Y_i \in \{0, 1\}$$

はグランドトゥルースラベルであり、

【数１１】

$$\hat{Y}_i \in [0, 1]$$

はｉ番目のサンプルに関する予測される提示の確率である。Ｌｉｎ　ｅｔ　ａｌ．［１０］は、Ｆｏｃａｌ　Ｌｏｓｓがデータの不均衡に対処するのに有効であることを明らかに

しており、Ｍｕｋｈｏｔｉ ｅｔ ａｌ．［１１］による最近の研究も、標準的な交差エントロピー損失と比較して、Ｆｏｃａｌ Ｌｏｓｓによってネットワークのキャリブレーションがより適切に行われることを示した。全体的な目的関数は、ペプチド－ＭＨＣ結合親和性に対するＭＳＥの変化形と、細胞表面へのペプチド提示に対する二値Ｆｏｃａｌ Ｌｏｓｓとの線形結合、$L = {}^{a}L_{BA-MSE} + (1-a)L_{p-FL}$である。

【０１７９】

トレーニング中に、例１．４で言及したネガティブサンプリング方法も適用した。これらを、提示タスク及び結合親和性タスクの両方について、各エポックの開始時に適用した。検証セットのネガティブサンプルを、最初のエポックの開始時に１回生成させ、トレーニングプロセス全体を通じて固定した。可能な場合には、それぞれのタスクから推測したもう一方のタスクの目標も利用した。サンプル重み法メカニズムも損失項毎に適用し、隣接領域変動の量に基づいてサンプルに異なる重み付けを付与し、目標が見つからないサンプルを覆い隠した（目標が不明な特定のタスクのサンプルのみを覆い隠した）。例１．５で説明した目標が推測されているサンプルは、両方のタスクに対して同一のサンプル重みを与えた。

【０１８０】

細胞表面へのペプチド提示の予測のために、二値交差エントロピー損失によって二値分類器のトレーニングを行う。L_{P-BCE}によって表される提示オブジェクティブ（ｐｒｅｓｅｎｔａｔｉｏｎ ｏｂｊｅｃｔｉｖｅ）は、

【数１２】

$$L_{\text{P-BCE}} = -\sum_{i} \left(Y_i \log \hat{Y}_i + (1 - Y_i) \log(1 - \hat{Y}_i)\right)$$

スキーム９

である。

【０１８１】

全体的な目的関数は、事前に規定した重み損失係数（ｗｅｉｇｈｔ ｌｏｓｓ ｃｏｅｆｆｉｃｉｅｎｔ）を用いた、上記のトレーニングオブジェクティブの線形結合：

【数１３】

$$L = \alpha L_{\text{P-BCE}} + \beta L_{\text{BA-MSE}} + \gamma L_{\text{A-MLM}}$$

スキーム１０

である。

【０１８２】

トレーニング中に、ランダムにサンプリングしたそれぞれの陰性サンプルについて、上記４つの方法のうちのいずれを使用するかについて、ネガティブサンプリング方法も適用した。陰性：陽性比のハイパーパラメータであるＮ比も定義し、これを使用して、それぞれの提示が陽性のサンプル毎に、サンプリングする陰性サンプルの数を決定した。サンプリングした陰性サンプル毎に１／Ｎ比のサンプル重みを割り当て、これにより、ＬＰ－ＢＣＥオブジェクティブ（ＬＰ－ＢＣＥ ｏｂｊｅｃｔｉｖｅ）が均衡の取れたデータ上でトレーニングを受けることが効果的に保証される（陽性及び陰性のトレーニングサンプルの全体的なサンプル重みが等しい）。

【０１８３】

さらに、主要なトレーニング段階の間に、覆い隠されたトークンが当該ペプチドの隣接領域からのみランダムに選択される、マスクされた言語モデリングの補助的オブジェクティブ（ａｕｘｉｌｉａｒｙ ｏｂｊｅｃｔｉｖｅ）を利用した。このオブジェクティブは天然のペプチド配列にのみ適用され、それ故に、ペプチド（例えばランダムにサンプリン

グされたアミノ酸配列）またはそのコンテキスト（例えばＨＲＶ陰性サンプル）のいずれかが「合成」である、特定のタイプのサンプリングされた陰性サンプルに関しては無視される。この補助的損失をＬ$_{A-MLM}$で表す。

【０１８４】

例２．４　キャリブレーション

Ｇｕｏ　ｅｔ　ａｌ．［１２］による最近の研究は、現代のニューラルネットワークには適切なキャリブレーションが行われていないことを明らかにしている。本発明者らのランキングロジックパイプラインにおいては、確率計算を使用し、予測された提示の確率に基づいて意思決定を行っていることから、キャリブレーションプロパティは本発明者らにとって非常に重要である。詳細には、ワクチン設計パイプラインにおいて、対照の６種のＨＬＡ対立遺伝子を考慮し、以下の確率計算を適用し、それぞれの単一のＨＬＡ対立遺伝子に対するモデルの予測に基づいて、当該対象の対立遺伝子に関する全体的な提示の確率を推定した。

【数１４】

$$P_{\text{提示}}(\text{ペプチド}|\text{対立遺伝子}_1,\cdots,\text{対立遺伝子}_6) = 1-\prod_{i=1}^{6}(1-P_{\text{提示}}(\text{ペプチド}|\text{対立遺伝子}_i))$$

スキーム４

【０１８５】

本発明者らの予測子に適切なキャリブレーションが行われていれば、この種の計算が向上することになる。したがって、低次多項式をキャリブレーション曲線に当て嵌めることにより、検証セット上でネットワークの提示の予測にさらにキャリブレーションを行った。単調増加関数を得るために、すべての多項式の係数が正になるように制約をかけた。このタスク用のＬａｓｓｏ線形回帰を使用した。このステップは、後に本発明者らの推論パイプラインに使用される、十分にキャリブレーションが行われた提示の確率を得るために重要であった。但し、このキャリブレーションステップは単調であることから、単一の対立遺伝子に関するペプチドのランキングには影響を与えない。本発明者らのワクチン設計パイプラインでは、免疫原性の確率推定に対する本発明者らの最良の代用値として、上記キャリブレーションを行った提示の予測を使用した。

【０１８６】

例２．５　モデルアーキテクチャ

Ａ．モデルアーキテクチャ１

このモデルのアーキテクチャは、３つの主要なコンポーネント、すなわち、配列処理、ペプチド－ＭＨＣ結合親和性予測、それに続く細胞表面へのペプチド提示予測で構成されていた。本発明者らのモデルへの入力は、上述の隣接領域、及び長さ４９アミノ酸のＨＬＡ対立遺伝子擬似配列を含むペプチド一次配列である。まず上記ペプチド配列を固定長ベクトルにコード化し、その後、共有のｄ次元アミノ酸埋め込み層を使用してすべてのアミノ酸配列をコード化した。次いで、埋め込まれた配列を平坦化し、各ペプチド、対立遺伝子、及び隣接領域のベクトル表現を生成させた。ペプチド－ＭＨＣ結合親和性予測コンポーネントについては、ペプチドとＨＬＡ対立遺伝子の表現を連結し、サイズが５１２と２５６の２つのｄｅｎｓｅ　ｌａｙｅｒ、ならびにそれぞれの後にＥｘｐｏｎｅｎｔｉａｌ　Ｌｉｎｅａｒ　Ｕｎｉｔ　（ＥＬＵ）活性化及びドロップアウト確率ｐ＝０．５のドロップアウト層を適用した。このコンポーネントの出力は「親和性表現」と呼ばれる。次に、追加のｄｅｎｓｅ　ｌａｙｅｒを使用して、予測された結合親和性ロジットを出力した。最後に、細胞表面へのペプチド提示予測コンポーネントについて、まずペプチド、隣接領域、及びＨＬＡ対立遺伝子の表現を単一のベクトルに連結させた。次に、ペプチド－ＭＨＣ結合親和性予測コンポーネント中と同様の構成をもつ一連の２つの全結合層を利用した

が、上記の「親和性表現」への出力も連結させた。このコンポーネントの出力は「提示表現」と呼ばれる。その後さらなる線形ｄｅｎｓｅ　ｌａｙｅｒを追加し、細胞表面へのペプチド提示の確率ロジットを予測した。

【０１８７】

Ｂ．モデルアーキテクチャ１

上記モデルは３つのコンポーネント、すなわち、１．配列埋め込み、２．Ｓｅｌｆ－ａｔｔｅｎｔｉｏｎ　Ｔｒａｎｓｆｏｒｍｅｒ層、及び３．予測ヘッドから構成される。

【０１８８】

１．配列埋め込み　－　上記モデルは入力として２つの配列、すなわち、１）アミノ酸配列（対立遺伝子擬似配列と隣接領域を含むペプチドとの連結）；２）各アミノ酸がどのセグメントに属するかをモデルに「通知」するセグメント識別子シーケンス（対立遺伝子、ペプチド、コンテキスト、及び特別なトークン）を受け取った。図１２は、トークン及びセグメント入力シーケンスの構成を示す。上記シーケンスの各トークンは、ａａ埋め込み、学習された位置の埋め込み、及びセグメント埋め込みの追加によって表される。

【数１５】

$$x = \mathrm{aa\_embed}(x) + \mathrm{pos\_embed}(x) + \mathrm{seg\_embed}(seg)$$

スキーム１１

【０１８９】

２．Ｓｅｌｆ－ａｔｔｅｎｔｉｏｎ　Ｔｒａｎｓｆｏｒｍｅｒ層　－　埋め込まれた配列を、図１３に示すように、それぞれがＭｕｌｔｉ－Ｈｅａｄ　Ｓｅｌｆ－ａｔｔｅｎｔｉｏｎ　ｍｏｄｕｌｅとそれに続くフィードフォワードモジュール（間にＧＥＬＵ活性化を有する２つの線形層で構成される）を含む１２の連続したＴｒａｎｓｆｏｒｍｅｒ層で処理した。層の正規化は各モジュールの開始時に適用され、割合ｐ＝０．１の残留ドロップアウトは各コンポーネントの終了時に残留接続の前に適用される。

【０１９０】

３．予測ヘッド　－　上記モデルは以下の３つの予測ヘッドを含まれていた。

【０１９１】

マスクされた言語モデリング：シーケンスの各位置における最終表現を、線形層（ＧＥＬＵ活性化及びそれに続く層正規化を含む）及び追加の学習バイアスを有するトークン埋め込み行列に重みを結合した追加の（トークン語彙サイズへの）線形射影で構成されるＬＭ予測ヘッドにフィードした。

【０１９２】

結合親和性：＜ＣＬＳ＞トークン位置における最終表現を、Ｌｉｎｅａｒ　＋　ＧＥＬＵ　＋　ＬａｙｅｒＮｏｒｍ　＋　Ｄｒｏｐｏｕｔ　＋　Ｌｉｎｅａｒで構成される予測ヘッドにフィードした。

【０１９３】

提示：単一の結合親和性ロジットと連結した、＜ＣＬＳ＞トークン位置における最終表現を、Ｌｉｎｅａｒ　＋　ＧＥＬＵ　＋　ＬａｙｅｒＮｏｒｍ　＋　Ｄｒｏｐｏｕｔ　＋　Ｌｉｎｅａｒで構成される予測ヘッドにフィードした。

【０１９４】

例２．６　モデルアンサンブル

上記のニューラルネットワークアーキテクチャを、２種類のモデル、すなわち、（１）すべてのトレーニングデータが含まれる汎対立遺伝子、及び（２）トレーニングデータがＨＬＡ対立遺伝子によって分割される対立遺伝子特異的のトレーニングに使用した。別個のＨＬＡ対立遺伝子特異的モデルに、各ＨＬＡ対立遺伝子毎に十分なトレーニングデータによるトレーニングを行う（データセット中に少なくとも１０００個の結合親和性サンプル及び１０００個の提示サンプルを有するという基準を利用した）。

【０１９５】
どちらのタイプも厳密に同一のアーキテクチャを有し、同様のトレーニングを受ける。上記２種類のモデルは、トレーニングデータ、すなわち、推論時にサポートを受ける対立遺伝子の一覧及びその重みの初期化のみが異なる。汎モデルはランダムに初期化を行う一方、対立遺伝子特異的モデルは最高のトレーニングを受けた汎モデルから微調整する。推論時には、トレーニングを受けたモデルのアンサンブルを利用し、これによりサンプルの対立遺伝子をサポートするすべてのモデルにわたる各サンプルの予測が平均化される。
【０１９６】
例２．７　モデルの選択
初期の実験では、多くの場合、モデルのサブセット、または単一のモデルの方が、多くのモデルのアンサンブルよりも優れたパフォーマンスを発揮することを示した。そこで、以下のモデル選択手順を開発した。１．所与のあらゆるモデルの構成（十分なトレーニングデータを有する各対立遺伝子に関する汎モデル及び対立遺伝子特異的モデル）について、異なる分割（ｆｏｌｄ）（トレーニング－検証分割）のみを使用した厳密なトレーニングセットアップで１０種のモデルのトレーニングを行う。トレーニングが完了した後に、検証免疫原性データ分割部分を使用した本発明者らの評価プロトコルに基づいて、モデル構成毎に１０種のトレーニング済みモデルの中から最もパフォーマンスの高い単一のモデルを選択する。対立遺伝子毎の階層モデル選択を適用し、該選択では、あらゆる対立遺伝子（パフォーマンス検証のための免疫原性データを有する）について、次の選択肢、すなわち、ａ）対立遺伝子特異的モデルのみを使用する；ｂ）汎モデルのみを使用する；ｃ）汎モデル＋対立遺伝子特異的モデルのアンサンブルを使用し、それらの予測を平均する；から、可能な限り最適な構成を選択する。
【０１９７】
例２．８　評価
モデルのパフォーマンスを評価するために、Ｔ細胞免疫原性データを利用して、所与のトレーニング済みモデルが、陽性の免疫原性ペプチド－ＨＬＡ対立遺伝子対に、どの程度適切に、陰性の非免疫原性の対よりも上位のランキングを行うように管理するかを確認した。詳細には、上位約２０位までのランキングがなされたペプチドに着目する。というのは、大まかに言って、これが所与のワクチンを製造するのに望ましいペプチドの量であるためである。免疫原性検証／テストセットを所与として、すべてのサンプルについてモデルの予測を抽出することができる。これを念頭に置いて、上位Ｋ位の項目、Ｐｒｅｃｉｓｉｏｎ＠Ｋ、ｎＤＣＧＫ、及びＲｅｃｉｐｒｏｃａｌ　Ｒａｎｋに焦点を当てた３つの共通のランキング指標を利用した。Ｏ’Ｄｏｎｎｅｌｌ　ｅｔ　ａｌ．［１３］などの以前の研究で使用された陽性的中率指標も利用した。各対立遺伝子について、免疫原性検証／テストセットにおいて、すべての対応するペプチドに、それらの予測される細胞表面へのペプチド提示の確率または予測される結合親和性スコアに基づいて別個にランキングを行い、各対立遺伝子のＰｒｅｃｉｓｉｏｎ＠Ｋ、ｎＤＣＧＫ、Ｒｅｃｉｐｒｏｃａｌ　Ｒａｎｋ、及び陽性的中率指標を算出した。
【０１９８】
各対立遺伝子についてこれらの指標を別個に計算した後に、加重平均を適用することにより、すべてのＨＬＡ対立遺伝子にわたってそれらを集計した。この場合、各対立遺伝子は米国の人口における頻度によって重み付けを行う。米国の４つの最大の倫理基準にかなった群の対立遺伝子頻度を使用した［１１］。これらのファイルによって導出された頻度に、米国人口の各民族の頻度によってさらに重み付けを行った［１２］。詳細には、ヨーロッパ系白人では０．５４、ヒスパニック系では０．２２、アフリカ系アメリカ人では０．１７、アジア系では０．０７を使用した。人口における対立遺伝子の頻度によって重み付けを行う動機は、上記指標が、対象であって、該対象に関してこの方法を理論的に役立てることができる可能性のある上記対象の割合を捉える（または少なくともより適切に相関付ける）ためである。免疫原性評価セットにおけるＨＬＡ対立遺伝子の人工に基づく頻度重みを図７にプロットする。最終的な指標であるＷｅｉｇｈｔｅｄ　Ｐｒｅｃｉｓｉｏ

ｎ＠Ｋ（ＷＰ＠Ｋ）、Ｗｅｉｇｈｔｅｄ ｎＤＣＧＫ（ＷｎＤＣＧＫ）、Ｗｅｉｇｈｔｅｄ－Ｒｅｃｉｐｒｏｃａｌ－Ｒａｎｋ（ＷＲＲ）、及びＷｅｉｇｈｔｅｄ Ｐｏｓｉｔｉｖｅ Ｐｒｅｄｉｃｔｉｖｅ Ｖａｌｕｅ（ＷＰＰＶ）は次式：

【数１６】

$$WP@K = \sum_i \tilde{f}_i \frac{\left| \text{陰性サンプル} \cap \text{上位1000位サンプル} \right|}{K_i}$$

$$WnDCG_K = \sum_i \tilde{f}_i \frac{DCG_K}{IDCG_K},$$

式中、

$$DCG_K = \sum_{k=1}^{K} \frac{2^{rel_k} - 1}{\log(k+1)},$$

である。

スキーム５

によって与えられる。ｒｅｌｋはｋ位にランキングされたアイテムであり、本発明者らのケースでは、これは二値免疫原性ラベルである。ＩＤＣＧＫは、基準アイテムの理想的なグラウンドトゥルースベースのランキングに対するＤＣＧｋスコアである。

【数１７】

$$WRR = \sum_i \frac{\tilde{f}_i}{\text{順位}_i}$$

スキーム６

【０１９９】

式中、ｒａｎｋｉは、ｉ番目の対立遺伝子のランキングされたペプチドにおける最初の陽性サンプルのランクである。

【数１８】

$$WPPV = \sum_i \tilde{f}_i \frac{\left| \text{陰性サンプル} \cap \text{上位} N_i \text{位サンプル} \right|}{N_i}$$

スキーム７

【０２００】

式中、Ｎｉはｉ番目の対立遺伝子に関する陽性サンプルの数である。すべての指標において、

【数１９】

$$\tilde{f}_i = f_i / \sum_j f_j$$

スキーム８

は、米国人口統計に基づいたＨＬＡ対立遺伝子頻度の重みであり、それらの値は［０，１］の範囲内にある（高いほど良好）。実際には、一部の対立遺伝子では、本発明者らの評価データ中の陽性サンプルが２０種未満である場合もあることから、ＷｎＤＣＧＫ指標に

Ｋ＝２０を使用し、ＷＰ＠ＫにはＫｉ＝ｍｉｎ（２０，｜陽性サンプル｜ｉ）を使用した。

【０２０１】

ＰＰＶ指標は、有益ではあるが、本発明者らが最も関心をもっていた領域である上位数種のペプチドに焦点を当てていない。Ｒｅｃｉｐｒｏｃａｌ Ｒａｎｋ指標は、ランキングされたペプチド内の最初の陽性アイテムのランクを使用し、ランクが高いほど高いスコアを与える。但し、この指標は、上位のランキングがなされたアイテムにおけるさらなる陽性ペプチドの有無を無視している。このことは理想的とは言えなかった。というのも、本発明者らが設計したワクチン中で陽性の可能性がある、可能な限り多くのサンプルを確実に得られるようにすることを望んでいたからである（提示されたすべてのペプチドが免疫原性であり、且つ所望の免疫応答を生み出すわけではないため）。Ｐｒｅｃｉｓｉｏｎ＠Ｋ指標は、単純に上位Ｋ位までにランキングされたアイテム内での陽性サンプルの割合を捉える。但し、上位Ｋ位以内の実際のランクを考慮してはいない（すなわち、１位のランキングを得た１種の陽性アイテムとＫ位のランキングを得た１種の陽性アイテムは全く同一のスコアを取得することになり、このことは理想的ではない）。ｎＤＣＧＫ指標では、上位Ｋ位以内の陽性サンプルの割合と、それらの対応するランクの両方が考慮される。理想的なランキングのＤＣＧによる正規化係数により、陽性サンプルの数が２０未満の場合にＫを制限する必要性も低減され、且つ［０，１］範囲内の値も生じる。但し、これらのスコアは、ｐｒｅｃｉｓｉｏｎ＠Ｋ指標と比較すると、解釈性／直観性の点でなおも幾分低い。

【０２０２】

例３．１　結果

別段の明示がない限り、０．００１の学習率と２５６のバッチサイズを使用して、ＡＤＡＭオプティマイザによりすべてのモデルのトレーニングを行った。損失項係数としてα＝０．５を設定し、両方の損失項に等しい重みを与えた。隠された検証セット上で評価した本発明者らの検証損失に、２０回のエポック後において改善が見られなかった場合には、早期停止を適用した。１ｅ－９という小さなＬＩ正則化係数と、ドロップ率が０．５のドロップアウトの両方を適用した。すべてのモデルに関して、ネガティブサンプリング方法及び目標の推測を適用した。対立遺伝子特異的モデルの場合、陽性サンプルの対立遺伝子を別のスーパータイプの対立遺伝子で置換することが望ましくなかった（特定の対立遺伝子のサンプルにのみ関心があるために）場合には、他のスーパータイプ（トレーニングを行う対象である特定の対立遺伝子のスーパータイプ（複数可）以外）の対立遺伝子を有する陽性サンプル、及びこのセットからランダムにサンプリングした陽性サンプルのリポジトリを、実行の順を逆にして着手しなかった（他の対立遺伝子すべてのデータのフィルタリングを行うまで）が、一方で「外部」対立遺伝子を、トレーニングを行っていた現在の対立遺伝子に置換した。最終的な予測子に関しては、以下のモデルタイプ、すなわち、１．パンモデル；及び２．対立遺伝子特異的モデル；のセットにトレーニングを行った。かかるモデルには、十分なトレーニングデータを有する対立遺伝子のサブセット上でのみトレーニングを行った。詳細には、少なくとも１０００種の結合親和性トレーニングサンプルと１０００種の提示トレーニングサンプルを有する対立遺伝子のみにトレーニングを行った。

【０２０３】

それぞれのタイプについて、１０種の類似のモデルに、異なるトレーニング検証分割部分上でトレーニングを行い、免疫原性検証セット上で最もパフォーマンスの高いモデルを選択した。階層モデルの選択を適用して、対立遺伝子毎に、推測の間にどのモデル（複数可）を使用するかを判定した。

【０２０４】

マルチタスクトレーニングが半発明者らの設定において有益であるかどうかを検証するために、損失重み係数（ｌｏｓｓ ｗｅｉｇｈｔ ｃｏｅｆｆｉｃｉｅｎｔ）αの効果を調査した。このハイパーパラメータを除いて同一の設定で、いくつかの汎モデルの３倍の

実験を実施し、免疫原性試験分割部分に対する平均ＷＰ＠Ｋを報告した。親和性予測によるランキングと提示予測によるランキングの両方によって取得した結果を、全ての効果を捉えるためにプロットした。図７は、α＝０及びα＝１（それぞれ、提示のみ及び親和性のみでのトレーニングに対応）により、両方のオブジェクティブの重み付けした組み合わせを使用した合同トレーニングに対応する中間領域での値と比較して、劣る結果が生じることを明確に示している。

【０２０５】

提示の確率が免疫原性推定の良好な代用値であるという本発明者らの仮説を検証するために、ビン分けされた提示の予測のヒストグラムを各ビンの免疫原性陽性サンプルの割合でプロットした。図８のＡに見られる正の傾きにより、$dP_{免疫原性}/dP_{提示}>0$、すなわち、平均で、ペプチドの提示の確率が増加すると、該ペプチドが免疫原性である可能性が高くなることが確認された。比較のために、結合親和性の予測に関する同等の挙動も調査し、図８のＢにおいて類似のパターンを観測した。また、本発明者らのモデルと他の最先端の予測子の間でのパフォーマンスの比較も行った。

【０２０６】

詳細には、Ｏ’Ｄｏｎｎｅｌｌ　ｅｔ　ａｌ．［１３］及びその以前のバージョンのＭＨＣｆｌｕｒｒｙ－２．０とパフォーマンスを比較した。上記比較は本発明者らの評価プロトコル及び指標を使用して実施し、報告した数値は免疫原性試験分割部分上で計算した。表１から明らかなように、本発明者らの最良の予測子は、ＭＨＣｆｌｕｒｒｙ－２．０予測子及び以前のバージョンのすべての変化形よりも有意に性能が優れていた。また、本発明者らの単一の汎モデルが、対立遺伝子特異的モデルのコレクションよりも性能が優れていることも明らかにすることができており、（上記コレクションがすべての対立遺伝子に対する予測をサポートしていないため、これは当然である）、さらに、階層モデルの選択を伴う、両者の組み合わせによって、パフォーマンスがさらに向上した。

【表１】

表１　提案されている免疫原性ベンチマークの評価結果。最良の予測子とＭＨＣｆｌｕｒｒｙ予測子のパフォーマンスとの比較。報告する指標は：Ｗｅｉｇｈｔｅｄ　Ｐｒｅｃｉｓｉｏｎ＠Ｋ　（ＷＰ＠Ｋ）、Ｗｅｉｇｈｔｅｄ　ｎＤＣＧＫ（ＷｎＤＣＧＫ）、Ｗｅｉｇｈｔｅｄ　Ｒｅｃｉｐｒｏｃａｌ　Ｒａｎｋ（ＷＲＲ）、及びＷｅｉｇｈｔｅｄ　Ｐｏｓｉｔｉｖｅ　Ｐｒｅｄｉｃｔｉｖｅ　Ｖａｌｕｅ（ＷＰＰＶ）であり、すべて高いほどより良好。

予測子		WP@K	$WnDCG_{20}$	WRR	WPPV
BigMHC-2.0	提示予測子	0.589	0.651	0.763	0.492
	親和性予測子	0.539	0.602	0.725	0.459
BigMHC-1.3	提示予測子	0.535	0.591	0.744	0.418
	親和性予測子	0.466	0.526	0.703	0.396
MHCflurry-2.0	提示予測子+隣接領域	0.421	0.501	0.627	0.343
	提示予測子(隣接領域なし)	0.424	0.5	0.659	0.348
	親和性予測子	0.468	0.547	0.69	0.383

【０２０７】

さらに、本発明者らの最良の予測子である「汎＋対立遺伝子特異的－提示予測子」と、ＭＨＣｆｌｕｒｒｙ－２．０の最良の予測子である「汎（＋ｍｓ）－親和性予測子」の間で、対立遺伝子毎のパフォーマンス（評価対象となる免疫原性データを有するすべての対立遺伝子について）を比較した。結果を表２に示す。

【表２－１】

表２　本明細書に記載の最良の予測子と、提案されている免疫原性のベンチマークであるＭＨＣｆｌｕｒｒｙ－２．０ベースラインの間での、対立遺伝子毎のパフォーマンスの比較。報告される指標は：Ｐｒｅｃｉｓｉｏｎ@Ｋ（Ｐ@Ｋ）、ｎＤＣＧ２０、Ｒｅｃｉｐｒｏｃａｌ　Ｒａｎｋ（ＲＲ）、及びＰｏｓｉｔｉｖｅ　Ｐｒｅｄｉｃｔｉｖｅ　Ｖａｌｕｅ（ＰＰＶ）であり、すべて高いほどより良好。

対立遺伝子	BigMHC-2.0				BigMHC-1.3				MHCflurry-2.0			
	P@K	*nDCG*$_{20}$	*RR*	*PPV*	*P@K*	*nDCG*$_{20}$	*RR*	*PPV*	*P@K*	*nDCG*$_{20}$	*RR*	*PPV*
HLA-	0.95	0.967	1.00	0.59	0.85	0.786	0.50	0.48	0.65	0.721	1.00	0.46
HLA-	0.00	0.256	0.07	0.00	0.00	0.387	0.20	0.00	0.00	0.500	0.33	0.00
HLA-	1.00	1.000	1.00	0.80	1.00	1.000	1.00	0.65	0.90	0.906	1.00	0.68
HLA-	0.33	0.296	0.50	0.33	0.33	0.358	0.33	0.33	0.00	0.257	0.12	0.00
HLA-	0.20	0.356	1.00	0.20	0.20	0.368	1.00	0.20	0.20	0.233	0.25	0.20
HLA-	0.42	0.655	1.00	0.42	0.42	0.661	1.00	0.42	0.42	0.726	1.00	0.42
HLA-	1.00	1.000	1.00	1.00	1.00	1.000	1.00	1.00	0.00	0.500	0.33	0.00
HLA-	0.00	0.000	0.04	0.00	0.00	0.000	0.03	0.00	0.00	0.500	0.33	0.00
HLA-	0.50	0.613	1.00	0.50	0.00	0.000	0.01	0.00	0.00	0.264	0.25	0.00
HLA-	0.85	0.876	1.00	0.57	0.75	0.814	1.00	0.40	0.65	0.704	1.00	0.44
HLA-	0.90	0.913	1.00	0.78	0.70	0.777	1.00	0.45	0.65	0.644	1.00	0.44
HLA-	0.00	0.000	0.02	0.00	0.00	0.000	0.02	0.00	0.00	0.000	0.02	0.00
HLA-	0.31	0.420	1.00	0.31	0.31	0.423	1.00	0.31	0.31	0.368	0.50	0.31
HLA-	0.45	0.443	0.50	0.46	0.50	0.584	1.00	0.41	0.50	0.542	1.00	0.35
HLA-	0.90	0.863	1.00	0.69	0.60	0.691	1.00	0.47	0.75	0.794	1.00	0.49
HLA-	0.61	0.774	1.00	0.61	0.61	0.777	1.00	0.61	0.53	0.666	1.00	0.53
HLA-	0.71	0.802	1.00	0.71	0.71	0.869	1.00	0.71	0.57	0.747	1.00	0.57
HLA-	0.00	0.237	0.20	0.00	0.50	0.613	1.00	0.50	0.50	0.613	1.00	0.50
HLA-	0.00	0.544	0.33	0.00	0.00	0.362	0.01	0.00	1.00	1.000	1.00	1.00
HLA-	1.00	1.000	1.00	1.00	0.00	0.500	0.33	0.00	1.00	1.000	1.00	1.00
HLA-	0.20	0.281	1.00	0.25	0.15	0.267	1.00	0.16	0.10	0.213	1.00	0.08
HLA-	0.66	0.735	1.00	0.66	0.60	0.584	0.50	0.60	0.60	0.735	1.00	0.60
HLA-	0.20	0.175	0.20	0.20	0.10	0.106	0.20	0.18	0.20	0.177	0.20	0.18
HLA-	0.54	0.630	1.00	0.54	0.54	0.678	1.00	0.54	0.54	0.690	1.00	0.54

【表２－２】

対立遺伝子	BigMHC-2.0				BigMHC-1.3				MHCflurry-2.0			
	P@K	*nDCG*$_2$	*RR*	*PPV*	*P@K*	*nDCG*$_2$	*RR*	*PPV*	*P@K*	*nDCG*$_2$	*RR*	*PPV*
HLA-	0.41	0.378	0.50	0.41	0.33	0.507	1.00	0.33	0.25	0.323	0.12	0.25
HLA-	0.27	0.442	1.00	0.27	0.27	0.493	1.00	0.27	0.36	0.511	1.00	0.36
HLA-	0.44	0.428	0.50	0.44	0.44	0.428	0.50	0.44	0.44	0.688	1.00	0.44
HLA-	0.50	0.616	0.50	0.50	0.50	0.628	1.00	0.50	0.50	0.697	0.50	0.50
HLA-	0.11	0.289	1.00	0.11	0.11	0.148	0.50	0.11	0.22	0.280	0.50	0.22
HLA-	0.00	0.000	0.02	0.00	0.00	0.000	0.03	0.00	0.00	0.000	0.03	0.00
HLA-	0.00	0.000	0.02	0.00	0.00	0.000	0.02	0.00	0.00	0.000	0.01	0.00
HLA-	0.70	0.776	1.00	0.60	0.45	0.551	1.00	0.48	0.55	0.628	1.00	0.52
HLA-	0.25	0.343	1.00	0.23	0.25	0.356	1.00	0.22	0.20	0.160	0.16	0.27
HLA-	0.33	0.587	1.00	0.33	0.00	0.202	0.25	0.00	0.33	0.366	0.33	0.33
HLA-	0.16	0.413	0.33	0.16	0.16	0.408	0.33	0.16	0.00	0.351	0.14	0.00
HLA-	0.00	0.117	0.06	0.00	0.00	0.107	0.05	0.00	0.33	0.576	1.00	0.33
HLA-	0.50	0.469	0.50	0.47	0.35	0.285	0.16	0.17	0.65	0.679	1.00	0.47
HLA-	0.95	0.967	1.00	0.69	0.90	0.935	1.00	0.60	0.75	0.700	0.50	0.54
HLA-	0.00	0.270	0.08	0.00	0.00	0.289	0.10	0.00	0.00	0.301	0.11	0.00
HLA-	0.90	0.912	1.00	0.54	0.75	0.796	1.00	0.42	0.60	0.623	1.00	0.36
HLA-	0.44	0.656	1.00	0.44	0.30	0.567	1.00	0.30	0.20	0.346	1.00	0.20
HLA-	0.50	0.867	1.00	0.50	0.50	0.855	1.00	0.50	0.50	0.817	1.00	0.50
HLA-	0.37	0.541	1.00	0.37	0.37	0.541	1.00	0.37	0.31	0.441	1.00	0.31
HLA-	0.60	0.723	1.00	0.60	0.60	0.723	1.00	0.60	0.40	0.648	1.00	0.40
HLA-	0.00	0.631	0.50	0.00	0.00	0.631	0.50	0.00	0.00	0.500	0.33	0.00
HLA-	0.45	0.432	0.33	0.34	0.35	0.325	0.25	0.29	0.30	0.288	0.33	0.25
HLA-	0.16	0.374	1.00	0.16	0.16	0.229	0.33	0.16	0.16	0.264	0.16	0.16
HLA-	0.50	0.635	1.00	0.52	0.55	0.609	1.00	0.52	0.45	0.524	1.00	0.42
HLA-	0.50	0.645	1.00	0.50	0.50	0.645	1.00	0.50	0.50	0.645	1.00	0.50
HLA-	0.00	0.000	0.02	0.00	0.00	0.000	0.00	0.00	0.00	0.000	0.00	0.00
HLA-	0.00	0.083	0.11	0.00	0.00	0.079	0.10	0.00	0.14	0.137	0.33	0.14
HLA-	0.00	0.167	0.16	0.00	0.00	0.141	0.11	0.00	0.00	0.246	0.09	0.00
HLA-	0.00	0.100	0.07	0.00	0.00	0.096	0.06	0.00	0.00	0.100	0.07	0.00
HLA-	0.55	0.662	1.00	0.37	0.55	0.670	1.00	0.38	0.60	0.702	1.00	0.36
HLA-	0.65	0.697	1.00	0.56	0.80	0.830	1.00	0.52	0.65	0.606	0.50	0.52
HLA-	0.00	0.231	0.05	0.00	0.00	0.256	0.07	0.00	0.00	0.000	0.01	0.00
HLA-	0.52	0.452	0.33	0.52	0.50	0.454	0.50	0.50	0.55	0.626	1.00	0.54
HLA-B*35:01	0.950	0.955	1.000	0.663	1.000	1.000	1.000	0.584	0.850	0.841	1.000	0.562
HLA-	0.10	0.104	0.33	0.14	0.15	0.132	0.20	0.17	0.20	0.158	0.14	0.11
HLA-	0.15	0.133	0.20	0.13	0.15	0.122	0.12	0.13	0.10	0.076	0.11	0.08
HLA-	0.00	0.289	0.10	0.00	0.00	0.315	0.12	0.00	0.00	0.356	0.16	0.00
HLA-	0.00	0.000	0.01	0.00	0.00	0.000	0.01	0.00	0.00	0.000	0.01	0.00
HLA-	0.16	0.454	1.00	0.16	0.16	0.268	0.50	0.16	0.16	0.371	0.50	0.16
HLA-	0.60	0.834	1.00	0.60	0.60	0.701	1.00	0.60	0.60	0.627	1.00	0.60
HLA-	0.10	0.281	1.00	0.10	0.10	0.220	1.00	0.10	0.10	0.220	1.00	0.10
HLA-	0.00	0.334	0.10	0.00	0.00	0.182	0.20	0.00	0.00	0.167	0.16	0.00
HLA-	0.70	0.789	1.00	0.50	0.70	0.772	1.00	0.45	0.60	0.665	1.00	0.41
HLA-	0.50	0.574	0.50	0.50	0.37	0.430	0.50	0.37	0.37	0.474	1.00	0.37
HLA-	0.50	0.911	1.00	0.50	0.50	0.888	1.00	0.50	0.50	0.839	1.00	0.50
HLA-	0.00	0.000	0.01	0.00	0.00	0.000	0.01	0.00	0.00	0.000	0.02	0.00
HLA-	0.45	0.375	0.33	0.37	0.40	0.397	0.50	0.37	0.35	0.280	0.20	0.34
HLA-	0.10	0.095	0.25	0.10	0.20	0.253	0.14	0.20	0.50	0.561	0.50	0.50
HLA-	0.21	0.312	1.00	0.21	0.26	0.360	1.00	0.26	0.36	0.484	1.00	0.36
HLA-	0.70	0.921	1.00	0.70	0.70	0.918	1.00	0.70	0.50	0.918	1.00	0.50
HLA-	0.58	0.606	0.33	0.58	0.58	0.690	1.00	0.58	0.58	0.651	0.50	0.58

【表２－３】

対立遺伝子	BigMHC-2.0				BigMHC-1.3				MHCflurry-2.0			
	P@K	*nDCG*$_2$	*RR*	*PPV*	*P@K*	*nDCG*$_2$	*RR*	*PPV*	*P@K*	*nDCG*$_2$	*RR*	*PPV*
HLA-	1.00	1.000	1.00	1.00	1.00	1.000	1.00	1.00	1.00	1.000	1.00	1.00
HLA-	0.45	0.652	1.00	0.45	0.45	0.495	0.33	0.45	0.36	0.412	0.33	0.36
HLA-	1.00	1.000	1.00	1.00	0.00	0.631	0.50	0.00	0.00	0.431	0.25	0.00
HLA-	0.00	0.000	0.01	0.00	0.00	0.000	0.04	0.00	0.50	0.387	0.50	0.50
HLA-	0.00	0.095	0.12	0.00	0.00	0.000	0.04	0.00	0.00	0.000	0.02	0.00
HLA-	0.00	0.343	0.25	0.00	0.33	0.402	0.33	0.33	0.00	0.343	0.25	0.00
HLA-	0.60	0.685	1.00	0.50	0.60	0.697	1.00	0.38	0.45	0.583	1.00	0.41
HLA-	0.35	0.481	1.00	0.35	0.20	0.250	0.50	0.20	0.20	0.275	1.00	0.20
HLA-	0.30	0.274	0.33	0.24	0.30	0.224	0.14	0.20	0.25	0.194	0.14	0.20
HLA-	0.40	0.534	0.33	0.40	0.40	0.637	0.33	0.40	0.60	0.667	0.33	0.60
HLA-B*55.01	0.12	0.160	0.50	0.12	0.12	0.219	0.50	0.12	0.12	0.160	0.50	0.12
HLA-	0.00	0.291	0.05	0.00	0.00	0.000	0.02	0.00	0.00	0.171	0.09	0.00
HLA-	0.70	0.712	1.00	0.50	0.70	0.714	1.00	0.45	0.70	0.745	1.00	0.40
HLA-	0.00	0.074	0.08	0.00	0.28	0.412	1.00	0.28	0.14	0.340	0.33	0.14
HLA-	0.21	0.278	0.33	0.21	0.20	0.352	0.50	0.20	0.26	0.300	0.16	0.26
HLA-	0.60	0.577	0.50	0.60	0.55	0.511	0.50	0.45	0.55	0.522	0.50	0.43
HLA-	0.50	0.613	1.00	0.50	0.00	0.136	0.10	0.00	0.00	0.000	0.01	0.00
HLA-	0.00	0.319	0.11	0.00	0.25	0.246	0.50	0.25	0.00	0.100	0.07	0.00
HLA-	0.00	0.441	0.25	0.00	0.00	0.482	0.25	0.00	0.00	0.403	0.16	0.00
HLA-	0.00	0.115	0.06	0.00	0.00	0.000	0.03	0.00	0.00	0.117	0.06	0.00
HLA-	0.25	0.332	1.00	0.23	0.15	0.161	0.33	0.19	0.20	0.146	0.10	0.19
HLA-	0.25	0.304	0.33	0.25	0.25	0.168	0.25	0.25	0.00	0.226	0.12	0.00
HLA-	0.00	0.000	0.04	0.00	0.00	0.000	0.04	0.00	0.14	0.275	1.00	0.14
HLA-	0.14	0.137	0.33	0.14	0.00	0.000	0.02	0.00	0.14	0.137	0.33	0.14
HLA-	0.00	0.060	0.05	0.00	0.12	0.185	0.25	0.12	0.25	0.306	0.33	0.25
HLA-	1.00	1.000	1.00	1.00	0.50	0.387	0.50	0.50	0.00	0.000	0.04	0.00
HLA-	0.16	0.108	0.16	0.16	0.00	0.000	0.02	0.00	0.00	0.095	0.12	0.00
HLA-	0.00	0.280	0.16	0.00	0.33	0.358	0.33	0.33	0.00	0.302	0.20	0.00
HLA-	0.00	0.000	0.01	0.00	0.00	0.000	0.01	0.00	0.00	0.000	0.01	0.00
HLA-	0.40	0.894	1.00	0.40	0.20	0.335	0.50	0.20	0.40	0.550	1.00	0.40
HLA-	0.16	0.151	0.33	0.16	0.16	0.303	1.00	0.16	0.00	0.143	0.06	0.00
HLA-	1.00	1.000	1.00	1.00	1.00	1.000	1.00	1.00	1.00	1.000	1.00	1.00
HLA-	0.28	0.318	0.33	0.28	0.28	0.296	0.25	0.28	0.14	0.183	0.20	0.14
HLA-	0.00	0.042	0.05	0.00	0.07	0.096	0.11	0.07	0.07	0.046	0.07	0.07
HLA-	0.15	0.222	1.00	0.14	0.20	0.180	0.33	0.19	0.15	0.125	0.16	0.14
HLA-	0.00	0.338	0.20	0.00	0.33	0.343	0.33	0.33	0.00	0.287	0.16	0.00
HLA-	0.00	0.000	0.01	0.00	0.00	0.000	0.00	0.00	0.00	0.000	0.01	0.00
HLA-	0.20	0.214	0.50	0.20	0.20	0.214	0.50	0.20	0.20	0.214	0.50	0.20
HLA-	0.25	0.390	1.00	0.25	0.25	0.344	0.50	0.25	0.00	0.139	0.16	0.00

【０２０８】

実施例で引用した参考文献

１．Ｖｉｔａ，Ｒ．；Ｍａｈａｊａｎ，Ｓ．；Ｏｖｅｒｔｏｎ，Ｊ．Ａ．；Ｄｈａｎｄａ，Ｓ．Ｋ．；Ｍａｒｔｉｎｉ，Ｓ．；Ｃａｎｔｒｅｌｌ，Ｊ．Ｒ．；Ｗｈｅｅｌｅｒ，Ｄ．Ｋ．；Ｓｅｔｔｅ，Ａ．；Ｐｅｔｅｒｓ，Ｂ．Ｔｈｅ ｉｍｍｕｎｅ ｅｐｉｔｏｐｅ ｄａｔａｂａｓｅ（ＩＥＤＢ）：２０１８ ｕｐｄａｔｅ．Ｎｕｃｌｅｉｃ ａｃｉｄｓ ｒｅｓｅａｒｃｈ ２０１９，４７，Ｄ３３９－Ｄ３４３．

２．Ｋｉｍ，Ｙ．；Ｓｉｄｎｅｙ，Ｊ．；Ｂｕｕｓ，Ｓ．；Ｓｅｔｔｅ，Ａ．；Ｎｉｅｌｓｅｎ，Ｍ．；Ｐｅｔｅｒｓ，Ｂ．Ｄａｔａｓｅｔ ｓｉｚｅ ａｎｄ ｃｏｍｐｏｓｉｔｉｏｎ ｉｍｐａｃｔ ｔｈｅ ｒｅｌｉａｂｉｌｉｔｙ ｏｆ ｐｅｒｆｏｒｍａｎｃｅ ｂｅｎｃｈｍａｒｋｓ ｆｏｒ ｐｅｐｔｉｄｅ－ＭＨＣ ｂｉｎｄｉｎｇ ｐｒｅｄｉｃｔｉｏｎｓ．ＢＭＣ ｂｉｏｉｎｆｏｒｍａｔｉｃｓ ２０１４，１５，２４１．

3. O'Donnell, T.J.;Rubinsteyn, A.;Bonsack, M.;Riemer, A.B.;Laserson, U.;Hammerbacher, J. MHCflurry: open-source class I MHC binding affinity prediction. Cell systems 2018, 7, 129-132.
4. Sarkizova, S.;Klaeger, S.;Le, P.M.;Li, L.W.;Oliveira, G.;Keshishian, H.;Hartigan, C.R.;Zhang, W.;Braun, D.A.;Ligon, K.L.;others. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. Nature Biotechnology 2020, 38, 199-209.
5. Shao, W.;Pedrioli, P.G.;Wolski, W.;Scurtescu, C.;Schmid, E.;Vizcaino, J. A.;Courcelles, M.;Schuster, H.;Kowalewski, D.;Marino, F.;others. The SysteMHC atlas project. Nucleic acids research 2018, 46, D1237-D1247.
6. Abelin, J.G.;Harjanto, D.;Malloy, M.;Suri, P.;Colson, T.;Goulding, S.P.;Creech, A.L.;Serrano, L.R.;Nasir, G.;Nasrallah, Y.;others. Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer epitope prediction. Immunity 2019, 51, 766-779.
7. Sidney, J.;Peters, B.;Frahm, N.;Brander, C.;Sette, A. HLA class I supertypes: a revised and updated classification. BMC immunology 2008, 9, 1.
8. O'Donnell, T.;Rubinsteyn, A.;Laserson, U. Improved predictive models for peptide presentation on MHC I. BioRxiv 2020.
9. Consortium, U. UniProt: a worldwide hub of protein knowledge. Nucleic acids research 2019, 47, D506-D515.
10. Lin, T.Y.;Goyal, P.;Girshick, R.;He, K.;Dollar, P. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980-2988.
11. Mukhoti, J.;Kulharia, V.;Sanyal, A.;Golodetz, S.;Torr, P.H.;Dokania, P.K. Calibrating Deep Neural Networks using Focal Loss. arXiv preprint arXiv:2002.09437 2020.
12. Guo, C.;Pleiss, G.;Sun, Y.;Weinberger, K.Q. On calibration of modem neural networks. arXiv preprint arXiv: 1706.04599 2017.
13. O'Donnell, T.J.;Rubinsteyn, A.;Laser

son, U. MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. Cell Systems 2020.
14. Fischer, W.;Perkins, S.;Theiler, J.;Bhattacharya, T.;Yusim, K.;Funkhouser, R.;Kuiken, C.;Haynes, B.;Letvin, N.L.;Walker, B.D.;others. Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. Nature medicine 2007, 13, 100-106.
15. Jurtz, V.;Paul, S.;Andreatta, M.;Marcatili, P.;Peters, B.;Nielsen, M. NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. The Journal of Immunology 2017, 199, 3360-3368.
16. Consortium, U. UniProt: a worldwide hub of protein knowledge. Nucleic acids research 2019, 47, D506-D515.

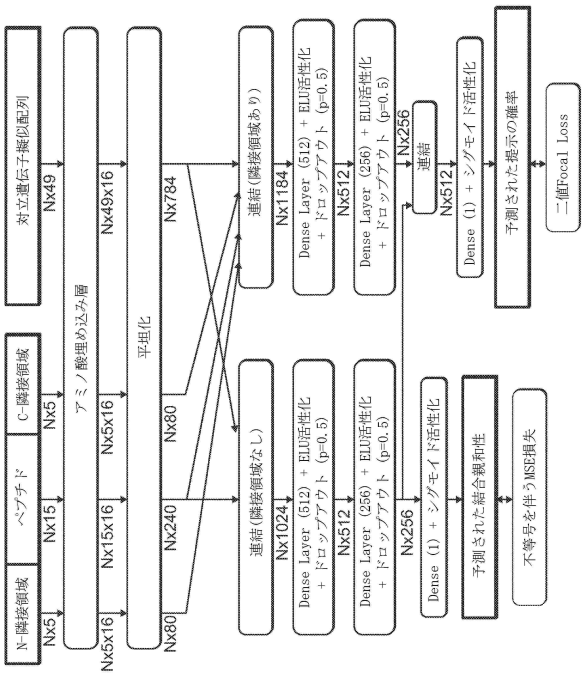
【0209】

等価物

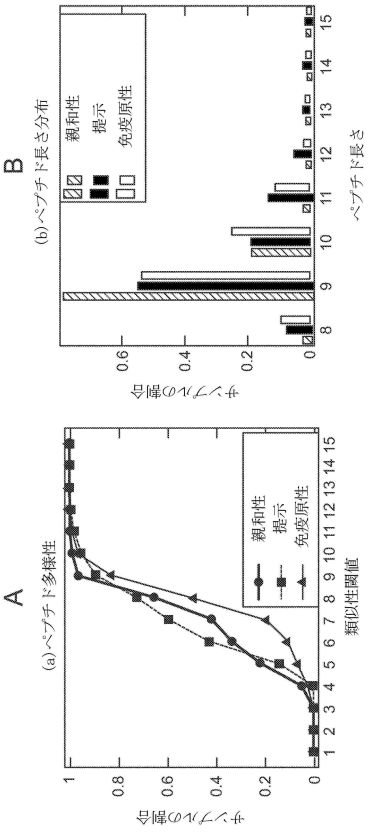
本明細書に記載の発明の方法の他の適切な修正及び適応は自明であること、及びこれらを、本開示または本実施形態の範囲から逸脱することなく、適切な等価物を使用して行うことができることは、当業者には容易に明らかであろう。ここまで、特定の組成物及び方法を詳細に説明してきたが、記載された実施例を参照することにより、上記組成物及び方法がより明確に理解されよう。これらの実施例は、説明のためにのみ紹介されており、限定を意図するものではない。

【図面】

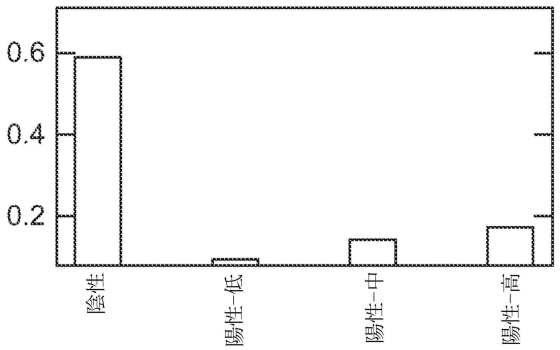
【図１】


対立遺伝子擬似配列
N-隣接領域
ペプチド
C-隣接領域
Nx5
Nx15
Nx5
Nx49
アミノ酸埋め込み層
Nx5x16
Nx15x16
Nx5x16
Nx49x16
平坦化
Nx80
Nx240
Nx80
Nx784
連結（隣接領域なし）
Nx1024
連結（隣接領域あり）
Nx1184
Dense Layer (512) + ELU活性化 + ドロップアウト (p=0.5)
Nx512
Dense Layer (512) + ELU活性化 + ドロップアウト (p=0.5)
Nx512
Dense Layer (256) + ELU活性化 + ドロップアウト (p=0.5)
Nx256
Dense (1) + シグモイド活性化
Nx256
連結
Nx512
Dense (1) + シグモイド活性化
予測された結合親和性
不等号を伴うMSE損失
予測された提示の確率
二値Focal Loss

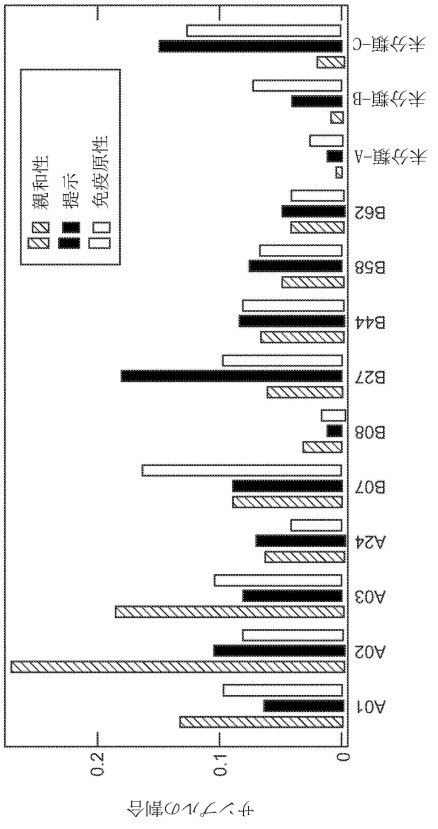
【図２】


A
(a)ペプチド多様性
サンプルの割合
類似性閾値
親和性
提示
免疫原性
B
(b)ペプチド長さ分布
サンプルの割合
ペプチド長さ

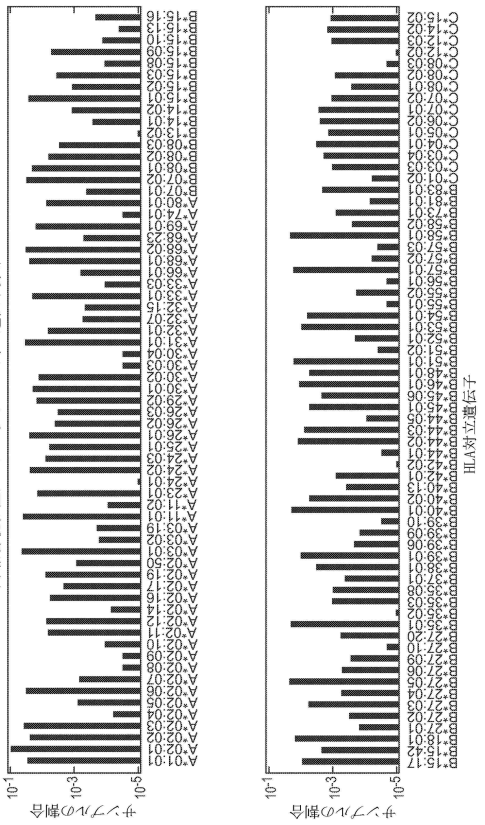
【図３】


0.6
0.4
0.2
陰性
陽性−低
陽性−中
陽性−高

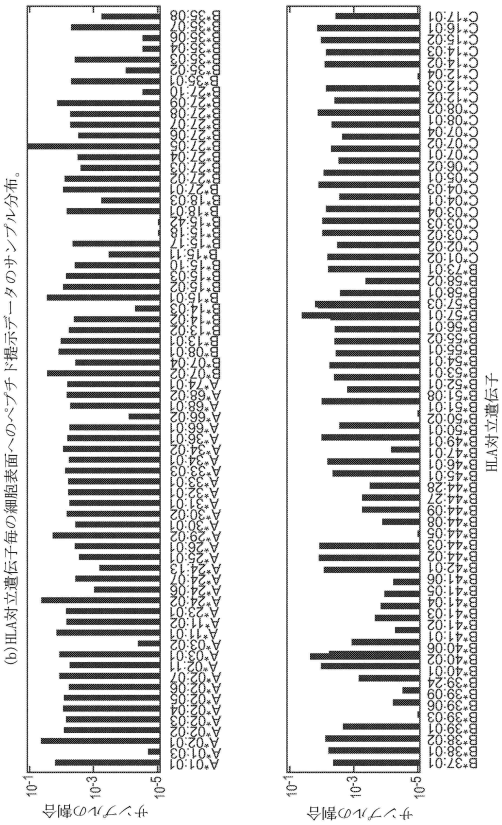
【図４】


サンプルの割合
親和性
提示
免疫原性
A01
A02
A03
A24
B07
B08
B27
B44
B58
B62
未分類-A
未分類-B
未分類-C

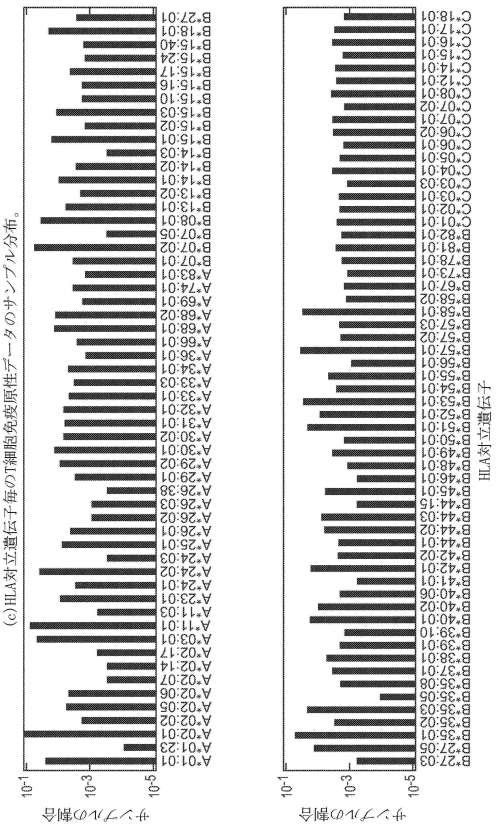
【図５Ａ】



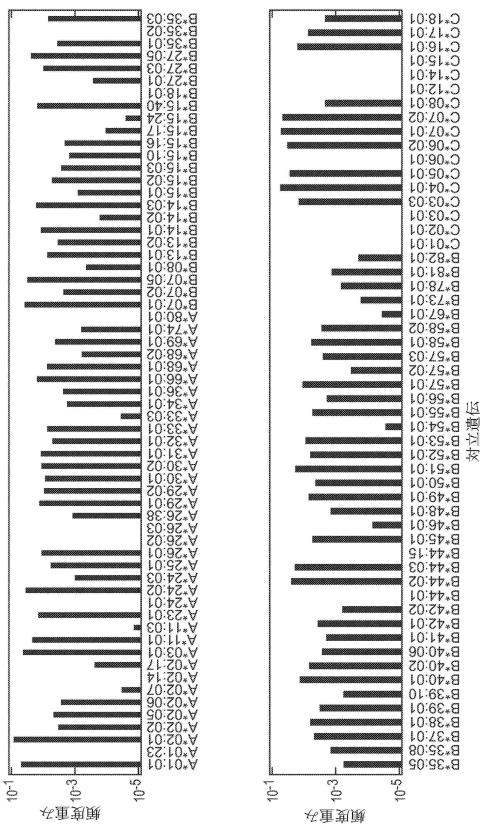
【図５Ｂ】



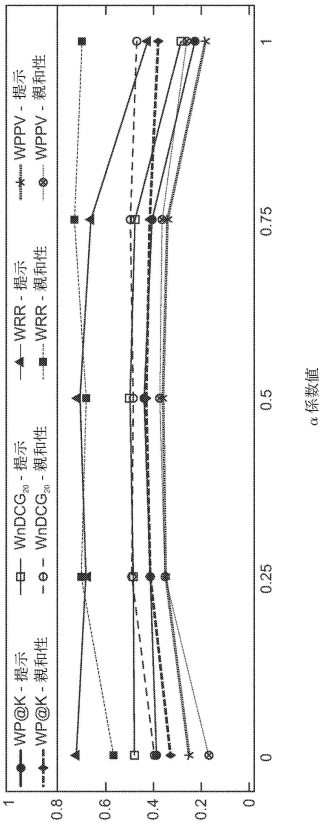
【図５Ｃ】



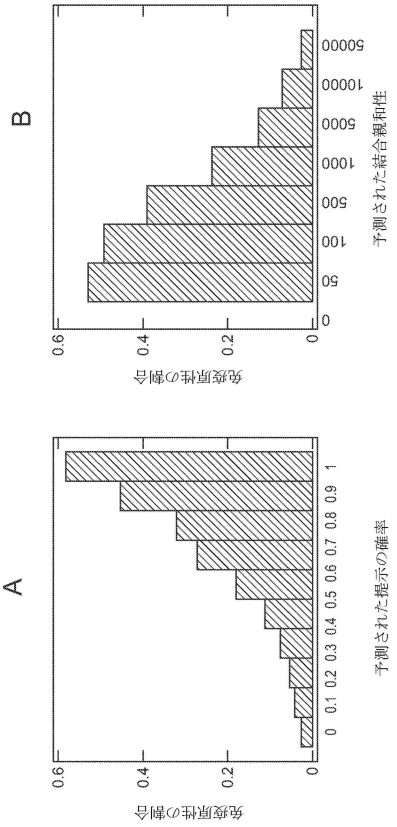
【図６】



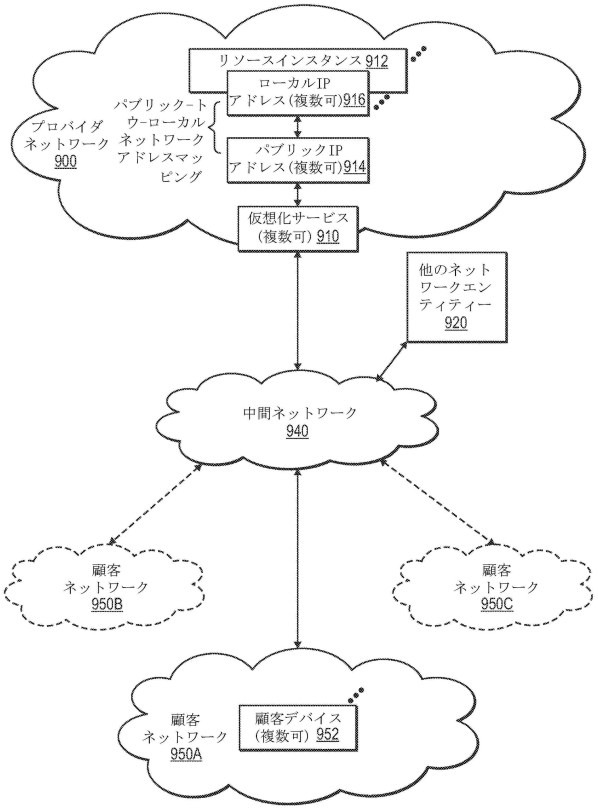
【図７】


WP@K - 提示
WP@K - 親和性
WnDCG@20 - 提示
WnDCG@20 - 親和性
WRR - 提示
WRR - 親和性
WPPV - 提示
WPPV - 親和性
α係数値
1
0.8
0.6
0.4
0.2
0
0.25
0.5
0.75

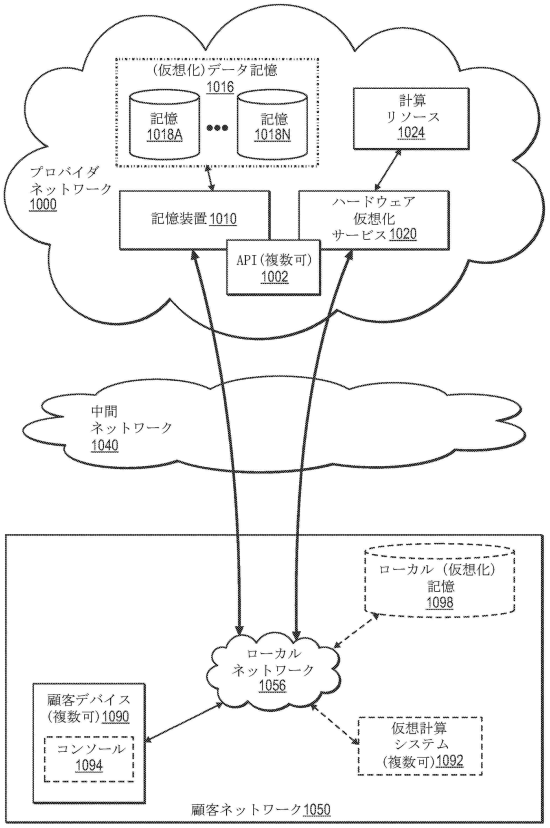
【図８】


B
免疫原性の割合
予測された結合親和性
50000
10000
5000
1000
500
100
50
0
0.6
0.4
0.2
A
予測された提示の確率
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

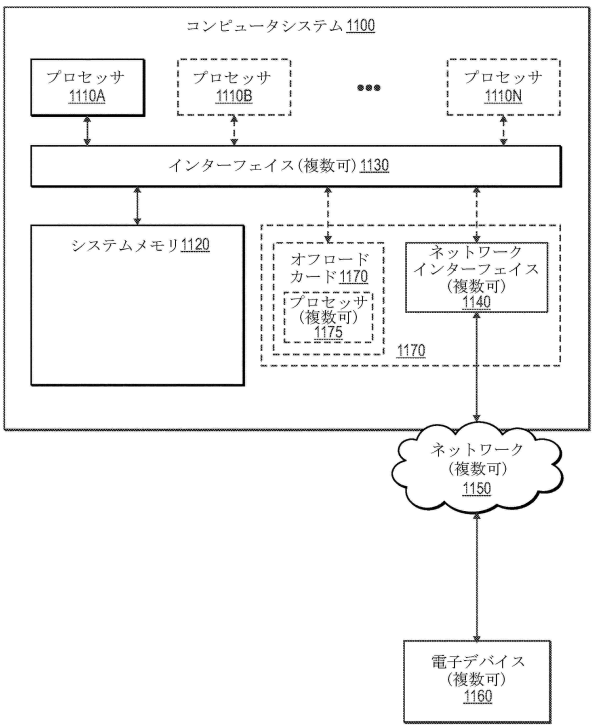
【図９】


リソースインスタンス 912
ローカルIP
アドレス(複数可)916
パブリック-トゥ-ローカル
ネットワーク
アドレスマッピング
パブリックIP
アドレス(複数可)914
プロバイダ
ネットワーク
900
仮想化サービス
(複数可) 910
他のネットワークエンティティー
920
中間ネットワーク
940
顧客
ネットワーク
950B
顧客
ネットワーク
950C
顧客
ネットワーク
950A
顧客デバイス
(複数可) 952

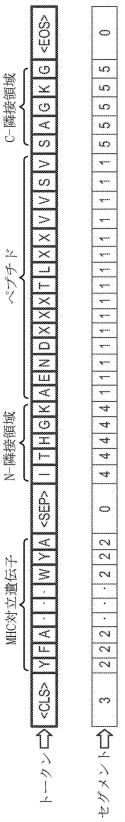
【図１０】


(仮想化)データ記憶
1016
記憶
1018A
記憶
1018N
計算
リソース
1024
プロバイダ
ネットワーク
1000
記憶装置1010
ハードウェア
仮想化
サービス1020
API(複数可)
1002
中間
ネットワーク
1040
ローカル(仮想化)
記憶
1098
ローカル
ネットワーク
1056
顧客デバイス
(複数可)1090
コンソール
1094
仮想計算
システム
(複数可)1092
顧客ネットワーク1050

【図１１】



【図１２】


<CLS> Y F A ・ ・ W Y A <SEP> I T H G K A E N D X X T L X V V S V S A G K G <EOS>
3 2 2 2 ・ ・ 2 2 2 0 4 4 4 4 1 1 1

【配列表】
20245056380000001.app

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2021/061399

A. CLASSIFICATION OF SUBJECT MATTER
INV. G16B15/30 G16B40/20 C12Q1/68 G06N20/00
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 C12Q G06N

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	O'DONNELL TIMOTHY J. ET AL: "MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing", CELL SYSTEMS, vol. 11, no. 1, 1 July 2020 (2020-07-01), pages 42-48.e7, XP055899634, US ISSN: 2405-4712, DOI: 10.1016/j.cels.2020.06.010 p. 42, "Introduction"; p. e4, "Peptide representation"; p. e4, "Allele representation"; p. e4, "Allele Multiple Alignment"; figures 1,2,3 -----	1-22

☐ 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	Date of mailing of the international search report
27 April 2022	10/05/2022

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	Heidrich, Alexander

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

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2021/061399

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:

   a. ☐ forming part of the international application as filed:

      ☐ in the form of an Annex C/ST.25 text file.

      ☐ on paper or in the form of an image file.

   b. ☐ furnished together with the international application under PCT Rule 13ter.1(a) for the purposes of international search only in the form of an Annex C/ST.25 text file.

   c. ☒ furnished subsequent to the international filing date for the purposes of international search only:

      ☒ in the form of an Annex C/ST.25 text file (Rule 13ter.1(a)).

      ☐ on paper or in the form of an image file (Rule 13ter.1(b) and Administrative Instructions, Section 713).

2. ☐ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that forming part of the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments:

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,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,IT,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,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,WS,ZA,ZM,ZW

（特許庁注：以下のものは登録商標）

１．ＰＹＴＨＯＮ

２．ＴＷＥＥＮ

(74)代理人　100162880
弁理士　上島 類
(72)発明者　ギル サデー
アメリカ合衆国 ワシントン シアトル テリー アヴェニュー ノース ４１０
(72)発明者　デイヴィッド ヘッカーマン
アメリカ合衆国 ワシントン シアトル テリー アヴェニュー ノース ４１０
(72)発明者　レイン クリストファー プライス
アメリカ合衆国 ワシントン シアトル テリー アヴェニュー ノース ４１０
(72)発明者　フランク ウィルヘルム シュミッツ
アメリカ合衆国 ワシントン シアトル テリー アヴェニュー ノース ４１０
(72)発明者　アンタ イマタ サフォ
アメリカ合衆国 ワシントン シアトル テリー アヴェニュー ノース ４１０
(72)発明者　ジャスリーン カウル グルーウォル
アメリカ合衆国 ワシントン シアトル テリー アヴェニュー ノース ４１０

Ｆターム（参考）　4B063　QA01 QA05 QA14 QA18 QQ02 QQ79 QQ96 QR48 QS10 QS15 QS33 QS39 QS40