



(12) 发明专利申请

(10) 申请公布号 CN 117136410 A

(43) 申请公布日 2023. 11. 28

(21) 申请号 202180095858.9

(22) 申请日 2021.12.01

(30) 优先权数据

63/139,074 2021.01.19 US

(85) PCT国际申请进入国家阶段日

2023.09.18

(86) PCT国际申请的申请数据

PCT/US2021/061399 2021.12.01

(87) PCT国际申请的公布数据

W02022/159176 EN 2022.07.28

(71) 申请人 亚马逊科技公司

地址 美国内华达州

(72) 发明人 吉尔·萨德 大卫·赫克曼

莱恩·克里斯托弗·普莱斯

弗兰克·威廉·施米茨

安塔·伊马塔·萨福

贾斯林·考尔·格鲁沃尔

(74) 专利代理机构 北京市京师律师事务所

11665

专利代理师 高晓丽

(51) Int.Cl.

G16B 15/30 (2006.01)

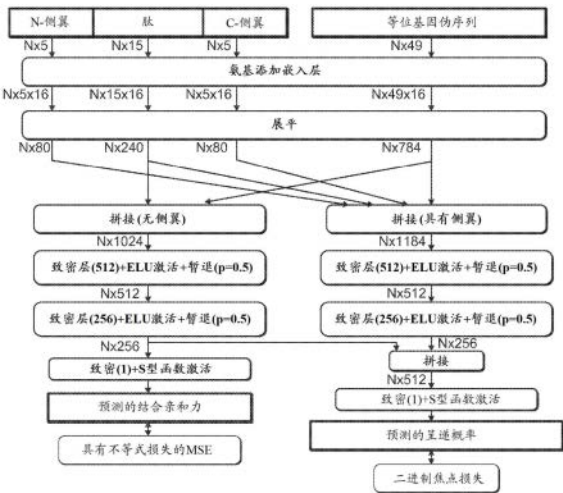
权利要求书2页 说明书38页 附图16页

(54) 发明名称

用于预测肿瘤特异性新抗原MHC I类或II类免疫原性的深度学习模型

(57) 摘要

本文公开了通过联合预测MHC I类或MHC II类结合亲和力并预测肿瘤特异性新抗原将由MHC I类或MHC II类蛋白在细胞表面上呈递的可能性来预测肿瘤特异性新抗原MHC I类或MHC II类免疫原性的方法。



1. 一种预测肿瘤特异性新抗原MHC I类免疫原性的方法,所述方法包括:

a) 获得肿瘤特异性新抗原的肽序列和所述肽序列的对应侧翼区;将所述肽序列和所述侧翼区编码到数值向量中,其中每个数值向量包含编码所述肿瘤特异性新抗原的所述肽的氨基酸残基和所述侧翼区的氨基酸残基以及氨基酸残基位置;

b) 获得HLA等位基因伪序列,其中所述HLA等位基因伪序列代表HLA等位基因;将所述HLA等位基因序列编码到对应的数值向量中;

c) 使用神经网络模型来联合预测肿瘤特异性新抗原MHC I类结合亲和力以及对于每个肿瘤特异性新抗原,对应肽将由MHC I类蛋白在细胞表面上呈递的数值概率;其中所述神经网络模型包括:

(i) 在训练数据集上训练所述神经网络模型以优化所述神经网络模型的性能;其中所述训练数据集包含肽-MHC I类亲和力测量数据集和细胞表面肽呈递数据集;

(ii) 输入层,所述输入层包含含有所述肿瘤特异性新抗原的所述肽序列和所述侧翼区的所述数值向量以及包含所述HLA等位基因伪序列层的所述数值向量;

(iii) 将包含所述肿瘤特异性新抗原的所述肽序列和所述侧翼区的所述数值向量和包含所述HLA等位基因伪序列的所述数值向量编码到氨基酸嵌入层中;

(iv) 将所述氨基酸嵌入层展平以产生所述肿瘤特异性新抗原的每个肽序列和所述肽序列侧翼区以及所述HLA等位基因伪序列的数值向量表示;

(v) 通过以下方式来预测所述肿瘤特异性新抗原MHC I类结合亲和力:拼接所述肿瘤特异性新抗原肽序列和所述HLA等位基因伪序列,应用一个或多个层和/或一个或多个激活函数,其中输出是代表所述肿瘤特异性新抗原MHC I类结合亲和力的数值得分;以及

(vi) 通过以下方式来预测所述肿瘤特异性新抗原将由MHC I类蛋白在细胞表面上呈递的概率:将所述感兴趣的肽序列、所述肽序列侧翼区和所述HLA等位基因伪序列拼接到单个数值向量中,应用一个或多个层和/或一个或多个激活函数,其中输出是肽将由MHC I类蛋白在细胞表面上呈递的数值概率;

其中所述肿瘤特异性新抗原MHC I类结合亲和力和所述肿瘤特异性新抗原将由所述MHC I类蛋白在细胞表面上呈递的所述数值概率是肿瘤特异性新抗原MHC I类免疫原性的指标。

2. 如权利要求1所述的方法,所述方法进一步包括通过以下方式验证所述神经网络模型:

(i) 将一个或多个排名度量应用于免疫原性验证数据集;

(ii) 基于所述肽的预测MHC I类结合亲和力以及所述肽将由MHC I类蛋白在细胞表面上呈递的数值概率,对所述免疫原性验证数据集中每个等位基因的肽进行排名;以及

(iii) 汇总所有等位基因的一个或多个排名度量。

3. 如权利要求2所述的方法,其中通过使用加权等位基因频率来汇总所述一个或多个排名度量。

4. 如前述权利要求中任一项所述的方法,其中所述神经网络模型是泛等位基因模型、等位基因特异性模型、超类型特异性模型或它们的组合。

5. 如前述权利要求中任一项所述的方法,其中所述HLA伪序列的长度为约30个氨基酸至约60个氨基酸。

6. 如前述权利要求中任一项所述的方法, 其中所述肿瘤特异性新抗原的所述肽序列的长度为约8个氨基酸至约15个氨基酸。

7. 如前述权利要求中任一项所述的方法, 其中所述侧翼区直接位于所述肿瘤特异性新抗原肽序列的左侧和/或直接位于所述肿瘤特异性新抗原肽序列的右侧。

8. 如前述权利要求中任一项所述的方法, 其中所述侧翼区的长度为约10个氨基酸。

9. 如权利要求8所述的方法, 其中直接位于所述肿瘤特异性新抗原左侧的所述侧翼区的长度为约5个氨基酸。

10. 如权利要求8所述的方法, 其中直接位于所述肿瘤特异性新抗原右侧的所述侧翼区的长度为约5个氨基酸。

11. 如权利要求1所述的方法, 所述方法进一步包括校准所述神经网络模型。

12. 如权利要求1所述的方法, 其中在阳性训练数据和阴性训练数据上训练所述神经网络模型。

13. 如权利要求12所述的方法, 其中所述阴性训练数据包含不具有所述肿瘤特异性新抗原MHC I类结合亲和力和/或不由所述MHC I类蛋白在细胞表面上呈递的肽。

14. 如前述权利要求中任一项所述的方法, 其中所述HLA等位基因是HLA A型、B型或C型。

15. 如前述权利要求中任一项所述的方法, 其中所述肿瘤特异性新抗原MHC I类免疫原性是CD8+T细胞免疫原性。

16. 如前述权利要求中任一项所述的方法, 其中针对免疫原性组合物选择一种或多种预测为MHC I类免疫原性的肿瘤特异性新抗原。

17. 如权利要求16所述的方法, 其中针对所述免疫原性组合物选择至少约20种肿瘤特异性新抗原。

18. 如前述权利要求中任一项所述的方法, 其中所述一个或多个层是完全连接层。

19. 如前述权利要求中任一项所述的方法, 其中所述一个或多个层是暂退层。

20. 如前述权利要求中任一项所述的方法, 其中所述一个或多个层和/或激活函数包括应用一个或多个完全连接层、应用暂退层以及应用激活函数。

21. 如权利要求11所述的方法, 其中用概率计算来校准所述神经网络模型。

22. 如权利要求11所述的方法, 其中所述概率计算估计受试者的等位基因的总体呈递概率。

用于预测肿瘤特异性新抗原MHC I类或II类免疫原性的深度学习模型

[0001] 本申请要求2021年1月19日提交的美国临时申请号63/139,074的权益,该美国临时申请的全部内容以引用方式并入本文。

1. 背景技术

[0002] 癌症是全世界死亡的主要原因,占有死亡的四分之一。Siegel等人,CA:ACancer Journal for Clinicians,68:7-30(2018)。2018年新癌症病例为1810万并且癌症相关死亡人数为960万。Bray等人,CA:ACancer Journal for Clinicians,68(6):394-424。有许多现有的标准治疗(standard of care)癌症疗法,包括消融技术(例如,外科手术和辐射)和化学技术(例如,化学治疗剂)。不幸的是,此类疗法常常与严重的风险、毒副作用和极高的成本以及不确定的功效相关联。

[0003] 癌症免疫疗法(例如,癌症疫苗)已出现作为一种有前途的癌症治疗方式。癌症免疫疗法的目标是控制并利用免疫系统来选择性破坏癌症,与此同时保持正常组织不受伤害。传统癌症疫苗通常靶向肿瘤相关抗原。肿瘤相关抗原通常存在于正常组织中,但在癌症中过表达。然而,由于这些抗原往往存在于正常组织中,所以免疫耐受可阻止免疫激活。几项靶向肿瘤相关抗原的临床试验未能证明与标准治疗疗法(standard of care treatment)相比持久的有益效应。Li等人,Ann Oncol.,28(增刊12):xii 11-xii 17(2017)。

[0004] 新抗原代表了癌症免疫疗法的有吸引力的靶标。新抗原是具有个体特异性的非自体同源蛋白质。新抗原源自肿瘤细胞基因组中的随机体细胞突变,并且不在正常细胞的表面上表达。出处相同。因为新抗原仅在肿瘤细胞上表达并且因此不会诱导中枢免疫耐受,所以靶向癌症新抗原的癌症疫苗具有潜在的优势,包括降低的中枢免疫耐受性和提高的安全性。出处相同。

[0005] 癌症的突变景观是复杂的,并且肿瘤突变对于每个个体受试者来说通常是独特的。通过测序检测到的大多数体细胞突变不会产生有效的新抗原。肿瘤DNA或肿瘤细胞中只有一小部分突变被以足以设计出可能有效的疫苗的准确性转录、翻译并加工成肿瘤特异性新抗原。另外,并非所有新抗原都是免疫原性的。事实上,自发识别内源性新抗原的T细胞的比例为约1%至2%。参见Paul等人,J.Immunol.,192,5831-5839(2013);Yewdell,Immunity,25,533-543(2006)。在大约1%的与MHC结合的新抗原中,只有大约50%的新抗原会被T细胞识别,并且只有30-40%的新抗原会被自然加工,从而使得能够进行肿瘤细胞杀伤。出处相同。

[0006] 当前的计算机模拟方法主要仅聚焦于建模哪些新抗原肽与MHC-I或MHC-II分子结合或预测哪些新抗原可能被肿瘤细胞加工成短肽并由MHC I/II类分子呈递。可得工具缺乏确定所呈递的肽中哪些肽具有免疫原性的预测准确性。因此,现有方法与低阳性预测值相关联。例如,在一项研究中,三名黑色素瘤患者各自用七种肽进行免疫,该七种肽的体外证实的MHC结合亲和力<500nM。Carreno等人,Science,348,803-808(2015)。在所测试的21种肽

中,只有9种肽诱导T细胞应答。出处相同。如果使用具有低阳性预测值的方法设计含有新抗原肽的个性化疫苗,则患者不太可能接受能够引发针对癌症的免疫应答的治疗性新抗原。此外,当前技术是耗时且费力的。

[0007] 因此,癌症患者中个性化新抗原的系统鉴定是成功开发个性化癌症疫苗的关键必要条件。因此,有效且准确地预测对个性化疫苗具有高阳性预测值的免疫原性新抗原候选物仍然是挑战。

2. 发明内容

[0008] 本公开涉及用于通过联合预测MHC I类或MHC II类结合亲和力并预测肿瘤特异性新抗原将由MHC I类或MHC II类蛋白在细胞表面上呈递的可能性来预测肿瘤特异性新抗原MHC I类或MHC II类免疫原性的新颖方法。所述方法准确地鉴定具有高预测价值的肿瘤特异性新抗原,所述肿瘤特异性新抗原可能被肿瘤细胞加工成结合受试者的MHC I类或MHC II类分子的肽,并且可能会接触T细胞受体并最终成为免疫原性的。这种方法对于鉴定将引发免疫应答的新抗原具有高预测准确性,这对于开发有效的个性化免疫原性组合物(例如,癌症疫苗)至关重要。这一直是使用现有方法时的障碍。

[0009] 此外,本文所述的方法优于金标准预测因子、MHCflurry-1.4结合亲和力预测因子以及MHCflurry-2.0预测因子。关于附加细节,请参见实施例小节。这些模型中的每个模型都是单独的预测因子。MHCflurry-1.4是等位基因特异性I类MHC结合预测因子(O'Donnell等人,Cell System,7:129-132(2018))。MHCflurry-2.0预测因子是MHC I类呈递肽的泛等位基因预测因子。

[0010] 本发明人已另外开发了免疫原性评估数据集来创建独特的基准,所述基准直接评估所述方法针对给定MHC I类等位基因从大型肽候选池中检索免疫原性肽的能力。这种能力是基于个性化免疫疗法的疫苗设计的重要组成部分。

[0011] 为了更清楚地描述用于预测肿瘤特异性新抗原MHC I类免疫原性的方法,图1提供了所述方法的示意性流程图。

[0012] 用于预测肿瘤特异性新抗原MHC I类或MHC II类免疫原性的方法开始于获得肿瘤特异性新抗原的肽序列和所述肽序列的对应侧翼区。所述侧翼区可以是直接位于所述肿瘤特异性新抗原肽的左侧或直接位于所述肿瘤特异性新抗原肽的右侧的氨基酸序列。例如,所述侧翼区可以是位于所述肿瘤特异性新抗原肽的C末端和/或N末端上的氨基酸序列。通常,侧翼区的长度可为约10个氨基酸。例如,直接位于所述肿瘤特异性新抗原的左侧的侧翼区的长度可为约5个氨基酸。例如,直接位于所述肿瘤特异性新抗原的右侧的侧翼区的长度可为约5个氨基酸。然后将所述肽序列和所述侧翼区编码到数值向量中。每个数值向量包含编码所述肿瘤特异性新抗原的肽和所述侧翼区的氨基酸残基以及氨基酸残基位置。获得代表HLA等位基因的HLA等位基因伪序列。HLA伪序列的长度可为至少约20至约100个氨基酸。优选地,HLA伪序列的长度为至少约30至60个氨基酸。将HLA等位基因伪序列编码到对应的数值向量中。HLA等位基因可为A型、B型、或C型、DQ型、DP型、或DR型。

[0013] 然后使用神经网络模型来联合预测肿瘤特异性新抗原MHC I类或MHC II类结合亲和力,以及对于每种感兴趣的肽,对应肽将由MHC I类或MHC II类蛋白在细胞表面上呈递的数值概率。神经网络模型可以是泛等位基因模型、等位基因特异性模型、超类型特异性模

型,或它们的组合。

[0014] 最初,在训练数据集上训练所述神经网络模型以优化所述神经网络模型的性能。所述训练数据集包括肽-MHC I类或MHC II类亲和力测量数据集和细胞表面肽呈递数据集。优选地,在阳性训练数据以及阴性训练数据上训练神经网络模型。阴性训练数据可包含不具有肿瘤特异性新抗原MHC I类或MHC II类结合亲和力和/或不由MHC I类或MHC II类蛋白在细胞表面上呈递的肽。

[0015] 模型输入层包括包含所述肿瘤特异性新抗原的肽序列和所述侧翼区的数值向量,以及包括含有所述HLA等位基因伪序列的数值向量在内的数值向量。接下来,将所述数值向量中的每个数值向量编码到氨基酸嵌入层中。然后,神经网络模型将氨基酸嵌入层展平以产生肿瘤特异性新抗原的每个肽序列和肽序列侧翼区以及HLA等位基因伪序列的数值向量表示。

[0016] 为了预测肽肿瘤特异性新抗原MHC I类或MHC II类结合亲和力,将肿瘤特异性新抗原肽序列和HLA等位基因伪序列拼接。所述模型进一步包括应用一个或多个层和/或一个或多个激活函数。例如,所述模型可包括应用一个或多个连接的层。例如,所述模型可包括应用暂退层(dropout layer)。例如,模型可包括应用激活函数。在实例中,模型可包括应用一个或多个连接的层、应用一个或多个暂退层、和/或应用激活函数。输出是代表肽配体MHC I类或MHC II类结合亲和力的数值得分。为了预测肿瘤特异性新抗原将由MHC I类或MHC II类蛋白在细胞表面上呈递的概率,将感兴趣的肽序列、肽序列侧翼区和HLA等位基因伪序列拼接到单个数值向量中。所预测的肽配体-MHC I类或MHC II类结合亲和力也被拼接。所述模型进一步包括应用一个或多个层和/或一个或多个激活函数。例如,所述模型可包括应用一个或多个连接的层。例如,所述模型可包括应用暂退层(dropout layer)。例如,所述模型可包括应用激活函数。所述模型可进一步包括应用一个或多个连接的层、应用暂退层、和/或应用激活函数。输出是肽将由MHC I类或MHC II类蛋白在细胞表面上呈递的数值概率。肿瘤特异性新抗原MHC I类或MHC II类结合亲和力以及肿瘤特异性新抗原将由MHC I类或MHC II类蛋白在细胞表面上呈递的数值概率是肿瘤特异性新抗原MHC I类免疫原性的指标。通常,MHC I类免疫原性是CD8+T细胞免疫原性。MHC II类免疫原性为CD4+免疫原性。

[0017] 所述方法可进一步包括通过以下方式来验证神经网络:将一个或多个排名度量应用于免疫原性验证数据集;基于免疫原性验证数据集中每个等位基因的肽的预测的MHC I类结合亲和力和所述肽将由MHC I类蛋白在细胞表面上呈递的数值概率对所述肽进行排名;以及汇总所有等位基因的排名度量。可以通过使用加权等位基因频率来汇总所述排名度量。所述方法可另外包括校准神经网络模型。所述神经网络模型可以用概率计算来校准。所述计算可以估计受试者的等位基因的总体呈递概率。

[0018] 可以选择被预测为MHC I类或MHC II类免疫原性的肿瘤特异性新抗原用于免疫原性组合。通常,可以选择约10至约20种肿瘤特异性新抗原用于免疫原性组合。

3.附图说明

[0019] 图1是模型架构图。模型输入:a)肽序列(加侧翼区);b)等位基因伪序列。模型输出:a)预测的结合亲和力;b)预测的呈递概率。用于训练模型的损失函数:a)用于结合亲和力预测的具有不等式损失的MSE;b)用于预测呈递概率的二进制焦点损失。

[0020] 图2A是描绘不具有相似性大于阈值的配偶体肽的肽的比率的图表。图2B是示出肽长度跨三个数据集(亲和性、呈递和免疫原性)的分布的图表。

[0021] 图3是示出肽-MHC结合亲和力标签分布的图表。y轴表示训练数据中属于每个箱的样本的比率。

[0022] 图4是示出数据集样本中HLA等位基因超类型的分布的图表。

[0023] 图5A至图5C是描绘每个HLA等位基因的样本分布的图表。图5A示出了肽MHC结合亲和力的样本分布。图5B示出了细胞表面肽呈递的样本分布。图5C示出了T细胞免疫原性的样本分布。

[0024] 图6是描绘基于美国群体中的等位基因频率的免疫原性等位基因频率权重的图表。

[0025] 图7是示出与不同 α 值的性能比较的图表,该性能比较确定了每个损失分量的权重。

[0026] 图8A和图8B是示出免疫原性率与预测的呈递概率(图8A)和预测的结合亲和力(图8B)之间的相关性的图表。这个实验是在免疫原性验证集上执行的,其中地面实况免疫原性标签是已知的并使用泛模型预测呈递和结合亲和力。

[0027] 图9示出了示例性提供商网络环境。

[0028] 图10是根据一些实施方案的向客户提供存储服务和硬件虚拟化的示例性提供商网络的框图。

[0029] 图11是示出示例性计算机系统的框图。

[0030] 图12示出了模型输入。所述模型接收两个序列,一个令牌序列和一个对应的分段序列。令牌序列是通过拼接<CLS>令牌、等位基因伪序列令牌、<SE>令牌、n-侧翼令牌、肽、c-侧翼令牌和<EOS>令牌而构成的。分段序列提供了指示对应令牌属于哪个分段的对应索引。

[0031] 图13是由多头自注意力模块之后是前馈模块(由2个线性层以及其间的GELU激活构成)构成的转换器层的示意图。在每个模块的开始处应用层归一化,并在残差连接之前在每个分量的末尾应用残差暂退。

4. 具体实施方式

[0032] 本公开涉及用于通过联合预测MHC I类或MHC II类结合亲和力并预测肿瘤特异性新抗原将由MHC I类或MHC II类蛋白在细胞表面上呈递的可能性来预测肿瘤特异性新抗原MHC I类或MHC II类免疫原性的新颖方法。所述新颖方法优选地用于预测肿瘤特异性新抗原MHC I类或MHC II类免疫原性。

[0033] 所述方法包括获得肿瘤特异性新抗原和代表HLA等位基因的HLA等位基因伪序列的测序数据。例如,外显子组、转录组和/或全基因组核苷酸测序可用于获得肿瘤特异性新抗原的测序数据和肽序列。所述方法可进一步包括将每种肿瘤特异性新抗原的肽序列和任选的侧翼区编码到对应的数值向量中。每个数值向量包括描述构成肽序列的氨基酸和氨基酸序列的位置的信息。

[0034] 所述方法还可包括将HLA伪序列编码到数值向量中。所述方法可包括将数值向量输入神经网络模型中以联合预测肿瘤特异性新抗原MHC I类或MHC II类结合亲和力,以及

针对每种肿瘤特异性新抗原,对应肽将由MHC I类或MHC II类蛋白在细胞表面上呈递的数值概率。这两种预测均可用于预测MHC I类或MHC II类免疫原性(例如,CD8+T细胞免疫原性或CD4+T细胞免疫原性)的指标。在将数值向量输入神经网络模型中之后,可以将数值向量转换成氨基酸嵌入层,然后可以将所述氨基酸嵌入层展平以产生肿瘤特异性新抗原的每个肽序列、任选地肽侧翼区和HLA等位基因伪序列的数值向量表示。

[0035] 接下来,神经网络模型可用于预测肿瘤特异性新抗原MHC I类或MHC II类结合亲和力和肿瘤特异性新抗原将由MHC I类或MHC II类蛋白在细胞表面上呈递的概率。这些预测可以通过拼接肿瘤特异性新抗原肽序列、HLA等位基因伪序列和任选的肽侧翼区来执行。然后可以应用一个或多个层和/或函数。例如,可以应用一个或多个完全连接的致密层。例如,可以应用一个或多个暂退层。例如,可以应用一个或多个激活函数。在实施方案中,可以应用一个或多个完全连接的致密层、一个或多个暂退层和/或一个或多个激活函数的组合。输出是代表肿瘤特异性MHC I类或MHC II类结合亲和力的数值得分和/或肽将由MHC I类或MHC II类蛋白在细胞表面上呈递的数值概率。这些预测值可以作为免疫原性的指标。然后可以选择免疫原性肿瘤特异性新抗原以包含在个性化免疫原性组合物中。

[0036] 本文公开的预测是基于训练数据集来鉴别的。所述训练数据集包括多个样本。训练数据集可包括肽MHC I类或MHC II类亲和力测量数据集和细胞表面肽呈递数据集。优选地,在阳性训练数据以及阴性训练数据上训练神经网络模型。阴性训练数据可包含不具有肿瘤特异性新抗原MHC I类或MHC II类结合亲和力和/或不由MHC I类或MHC II类蛋白在细胞表面上呈递的肽。

[0037] 可以选择被预测为MHC I类或MHC II类免疫原性的新抗原以包含在免疫原性组合物中。通常,可以选择约10至约20种肿瘤特异性新抗原用于免疫原性组合物。例如,大致免疫原性组合物可包含约5种、约6种、约7种、约8种、约9种、约10种、约11种、约12种、约13种、约14种、约15种、约16种、约17种、约18种、约19种、约20种、约21种、约22种、约23种、约24种、或约25种肿瘤特异性新抗原。

[0038] I. 定义

[0039] 本公开中引用的所有出版物和专利均全文以引用方式并入。如果以引用方式并入的材料与本说明书相矛盾或不一致,则本说明书将取代任何此类材料。本文对任何参考文献的引用并不承认这些参考文献是本公开的现有技术。整个说明书和权利要求书中使用了与本说明书的各方面相关的各种术语。除非另有说明,否则应给予此类术语在本领域中的普通含义。其他具体定义的术语将以与本文所提供的定义一致的方式解释。

[0040] 如本文所用,除非上下文另有明确指示,否则单数形式“一”、“一个(种)”和“该”包括复数形式。除非另外具体指出,否则术语“包括(含)”、“诸如”等旨在传达包括但不限于。

[0041] 如本文所用,术语“癌症”是指受试者中细胞群的特征在于不受控制的增殖、永生性、转移性潜力、快速生长和增殖速率和/或某些形态学特征的生理状况。通常癌症可以是肿瘤或肿块的形式,但可以单独存在于受试者体内,或者可以作为独立细胞(诸如白血病或淋巴瘤细胞)在血流中循环。术语癌症包括所有类型的癌症和转移瘤,包括血液恶性肿瘤、实体瘤、肉瘤、癌和其他实体和非实体肿瘤。癌症的示例包括但不限于癌、淋巴瘤、母细胞瘤、肉瘤和白血病。此类癌症的更具体示例包括鳞状细胞癌、小细胞肺癌、非小细胞肺癌、肺腺癌、肺鳞癌、腹膜癌、肝细胞癌、胃肠癌、胰腺癌、成胶质细胞瘤、宫颈癌、卵巢癌、肝癌

(liver cancer)、膀胱癌、肝细胞瘤 (hepatoma)、乳腺癌 (例如, 三阴性乳腺癌、激素受体阳性乳腺癌)、骨肉瘤、黑素瘤、结肠癌、结直肠癌、子宫内膜癌 (例如, 浆液性) 或子宫癌、唾液腺癌、肾癌、肝癌、前列腺癌、外阴癌、甲状腺癌、肝细胞瘤和各种类型的头颈癌。三阴性乳腺癌是指对雌激素受体 (ER)、黄体酮受体 (PR) 和 Her2/neu 基因表达呈阴性的乳腺癌。激素受体阳性乳腺癌是指对以下中的至少一者呈阳性: ER 或 PR, 并且对 Her2/neu (HER2) 呈阴性的乳腺癌。

[0042] 如本文所用的术语“新抗原”是指具有至少一种改变的抗原, 所述至少一种改变例如经由肿瘤细胞中的突变或特异于肿瘤细胞的翻译后修饰使所述抗原不同于对应的亲本抗原。突变可包括移码、插入/缺失 (indel)、错义或无义取代、剪接位点改变、基因组重排或基因融合、或产生新抗原的任何基因组表达改变。突变可包括剪接突变。特异于肿瘤细胞的翻译后修饰可包括异常磷酸化。特异于肿瘤细胞的翻译后修饰还可包括蛋白酶体生成的剪接抗原。参见 Lipe 等人, *Science*, 354 (6310): 354:358 (2016)。通常, 点突变占肿瘤中突变的约 95%, 并且其余的为插入/缺失和移码突变。参见 Snyder 等人, *N Engl J Med.*, 371:2189-2199 (2014)。

[0043] 如本文所用, 术语“肿瘤特异性新抗原”是存在于受试者的肿瘤细胞或组织中, 但不存在于受试者的正常细胞或组织中的新抗原。

[0044] 如本文所用, 如本文所用的术语“免疫原性”是指引发免疫应答 (例如, T 细胞应答、B 细胞应答或两者) 的能力。

[0045] 如本文所用, 术语“HLA 等位基因伪序列”是指通过算法生成以代表 HLA 等位基因氨基酸序列的氨基酸序列。

[0046] 如本文所用的术语“受试者”是指任何动物, 诸如任何哺乳动物, 包括但不限于人、非人灵长类动物、啮齿类动物等。在一些实施方案中, 哺乳动物是小鼠。在一些实施方案中, 哺乳动物是人类。

[0047] 如本文所用的术语“肿瘤细胞”是指作为癌细胞或来源于癌细胞的任何细胞。术语“肿瘤细胞”还可以指表现出癌症样性质的细胞, 所述癌症样性质为例如不受控制的繁殖、对抗生长信号的抗性、转移能力、以及经历程序性细胞死亡的能力的丧失。

[0048] 如本文所用, 术语“神经网络”是指用于分类或回归的机器学习模型, 所述机器学习模型由多层线性变换之后是通常经由随机梯度下降和反向传播训练的逐元素非线性组成。

[0049] 本文未直接定义的任何术语应被理解为具有如本发明的领域内所理解的通常与所述术语相关联的含义。本文未直接定义的任何术语应被理解为具有如本发明的领域内所理解的通常与它们相关联的含义。本文讨论了某些术语以向从业人员提供描述本发明的各方面的组合物、装置、方法等以及如何制造或使用它们的附加指导。应当理解的是, 同一件事可以用多于一种方式来叙述。因此, 替代语言和同义词可用于本文所讨论的术语中的任何一个或多个术语。在本文中是否详细说明或讨论术语并不重要。提供了一些同义词或可替代的方法、材料等。除非明确说明, 否则一个或几个同义词或等同物的叙述并不排除使用其他同义词或等同物。示例 (包括术语的示例) 的使用仅出于说明性目的, 并且不限制本文中的本发明的各方面的范围和含义。

[0050] 本文提供了方法的附加描述和方法实践的指导。为了便于呈递, 提供了关于通过

联合预测MHC I类结合亲和力和预测肿瘤特异性新抗原将由MHC I类蛋白在细胞表面上呈递的可能性来预测MHC I类免疫原性的优选方面的进一步细节和指导。进一步的细节和指导还涉及预测MHC II类免疫原性。

[0051] II. 训练

[0052] 神经网络模型可包括在训练数据集上训练神经网络模型以优化神经网络模型的性能,使得神经网络可以预测肿瘤特异性新抗原MHC I类或MHC II类结合亲和力以及肿瘤特异性新抗原将由MHC I类或MHC II类蛋白在细胞表面上呈递的概率。

[0053] 本文所述的方法中使用的训练数据集包括多个样本。训练数据可包括各种数据。训练数据集可包括肽MHC I类或MHC II类亲和力测量数据集和细胞表面肽呈递数据集,以及任选的免疫原性数据集。训练数据集可包括人类鼻病毒数据。阴性样本可用于免疫原性评估。训练数据集可用于训练一个或多个神经网络模型。在实施方案中,可以训练一个或多个神经网络模型。例如,可以训练至少2个或更多个神经网络。可以训练至少约3个、4个、5个、6个、7个、8个、9个、10个或更多个神经网络。

[0054] 包含肽MHC I类或MHC II类亲和力测量数据的数据集可包含实验测量的对特定MHC I类等位基因或MHC II类等位基因的结合亲和力肽。所述数据集可以从一个或多个数据源,诸如公开可得的数据源获得。例如,所述数据集可以从免疫表位数据库(“IEDB”, iedb.org)获得。所述训练数据集可以基于一个或多个数据源进一步扩充。可以针对本文所公开的方法进一步精选训练数据集。例如,训练数据可包含肽与相关联的MHC分子中的每个相关联的MHC分子之间的结合亲和力预测。数据集中实验测量的结合亲和力可以是定量的(与不等式“=”相关联)、定性的(与不等式“<”或“>”相关联),或它们的组合。定量数据可以包括IC₅₀ mM值。定性数据集可表示为阳性高(例如,结合亲和力<100nm)、阳性中等(例如,结合亲和力<1,000nm)、阳性低(例如,结合亲和力<5,000nm),或阴性(例如,结合亲和力>5,000nm)。包含MHC I类或MHC II类亲和力测量数据集的训练数据集可包含从MHC洗脱并通过质谱法鉴定的肽。

[0055] 可以进一步精选MHC I类亲和力测量数据集以保留特定MHC I类肽等位基因的预测结合亲和力的子集。例如,可以保留HLA-A、HLA-B和/或HLA-C等位基因的条目。例如,可以保留特定长度的肽。可以保留长度为至少约5个氨基酸至约20个氨基酸的肽。肽的氨基酸长度可为约5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个或20个氨基酸。训练集中的肽可以具有相同或不同的长度,并且可以取决于MHC等位基因的类型而变化。优选地,肽的长度为约5个至约15个氨基酸。可以剔除含有翻译后修饰或非典型氨基酸的肽。

[0056] 可以进一步精选MHC II类亲和力测量数据集以保留特定MHC II类肽等位基因的预测结合亲和力的子集。例如,可以保留HLA-DP、HLA-DQ和/或HLA-DR等位基因的条目。例如,可以保留特定长度的肽。可以保留长度为至少约5个氨基酸至约40个氨基酸的肽。肽的氨基酸长度可为约5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个、21个、22个、23个、24个、25个、26个、27个、28个、29个、30个、31个、32个、33个、34个、35个、36个、37个、38个、39个或40个氨基酸。训练集中的肽可以具有相同或不同的长度,并且可以取决于MHC等位基因的类型而变化。通常,肽的长度为约13个至约35个氨基酸。可以剔除含有翻译后修饰或非典型氨基酸的肽。

[0057] MHC I类或MHC II类亲和力测量数据可以作为回归模型提供。特别地,可以使用损失函数。示例性损失函数包括交叉熵损失函数、均方误差、胡贝尔损失、Kullback-Leibler、MAE (L1)、MAE (L3)、似然函数和合页损失。特别地,可以使用均方损失函数的变体。均方损失函数可由 L_{BA-MSE} 表示,其中测量与(>)或(<)相关联,仅在不等式被违反时才有助于损失,用于处理数据集中的定量和定性肽-MHC结合亲和力测量两者。可以采用的表达式为:

$$[0058] \quad L_{BA-MSE} = \frac{1}{N} \sum_{i=1}^N l(Y_i, \hat{Y}_i)$$

$$[0059] \quad l(\hat{Y}_i, Y_i) = \begin{cases} (Y_i - \hat{Y}_i)^2 & , \text{如果测量结果} i \text{的不等式是}(=) \\ \max(\hat{Y}_i - Y_i, 0)^2 & , \text{如果测量结果} i \text{的不等式是}(<) \\ \max(Y_i - \hat{Y}_i, 0)^2 & , \text{如果测量结果} i \text{的不等式是}(>) \end{cases}$$

[0060] 方案1

[0061] Y_i 和 $\hat{Y}_i$ 代表第*i*个样本的目标和预测的肽-MHC结合亲和力值。亲和力目标可以在训练前进行转换。 $[0, 50000]$ 范围中的IC50 nM值至 $[0, 1]$ 范围内的目标值。可以采用以下函数来转换IC50 nM值:

$$[0062] \quad Y_i = 1 - \frac{\log IC50_i}{\log 50000}$$

[0063] 方案2

[0064] MHC I类或MHC II类亲和力测量数据集可包含至少约5,000个、约10,000个、约15,000个、约20,000个、约25,000个、约30,000个、约35,000个、约40,000个、约45,000个、约50,000个、约60,000个、约70,000个、约80,000个、约90,000个、约100,000个、约150,000个、约200,000个、约250,000个、约300,000个、约350,000个、约400,000个、约450,000个、约500,000个、约550,000个、约600,000个、约650,000个、约700,000个、约750,000个、约800,000个、约850,000个、约900,000个、约950,000个、约1,000,000个、约1,250,000个、约1,500,000个、约1,750,000个、约2,000,000个或更多个肽与MHC I类或MHC II类肽等位基因的结合亲和力的测量结果。通常,MHC I类或MHC II类亲和力测量数据集包含至少约20,000个独特的肽。

[0065] 细胞表面肽呈递数据集可以包含已知经由HLA分子呈递的肽。例如,可以通过肽洗脱实验或通过质谱数据来确定细胞表面肽。细胞表面肽呈递数据集可以从一个或多个数据源,诸如公开可得的数据源获得。例如,免疫表位数据库(“IEDB”, iedb.org)或SystemMHC项目中生成的肽可能是有用的数据源。细胞表面肽呈递数据集可以进一步实验地产生。例如,可以通过稀释来自表达HLA肽的细胞系的肽并通过质谱法分析所述肽来制备肽。所述训练数据集可以基于一个或多个数据源进一步扩充。

[0066] 这些训练数据集通常是针对本文公开的方法而精选的。所述肽序列通常表示为字符串,其中每个字符代表氨基酸。肽序列可以被转换成数值向量,所述数值向量包括描述肽的氨基酸和所述氨基酸的位置的信息。数值向量可以是二进制分类。例如,具有 k^i 个氨基酸的肽序列 p^i 由20个氨基酸(20- k)的行向量表示,其中对应于所述肽序列的特定位置处的氨基酸的字母表的单个要素的值为1。其余要素的值将为0。作为示例,对于氨基酸字母表A、C、D、E、F、G、H、I、K、L、M、N、P、Q、R、S、T、V、W和Y,3个氨基酸的肽序列AFP可以由60个要素的行向

量表示,并且 $p^i=1\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0$
 $0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0$ 。当训练数据集包含不同长度的氨基酸序列时,数值向量可包含填充字符以将肽编码成相等长度。填充字符可以应用在肽序列的左侧或肽序列的右侧。本领域技术人员将认识到可以应用其他类型的分类系统。

[0067] 可以对细胞表面肽呈递数据集采用损失函数。示例性损失函数包括交叉熵损失函数、均方误差、胡贝尔损失、Kullback-Leibler、MAE (L1)、MAE (L3)、似然函数和合页损失。特别地,可以由细胞表面肽呈递数据集采用交叉熵损失函数。在特定实施方案中,可以使用焦点损失二进制分类。可以采用焦点损失二进制分类来减少数据集中的不平衡。焦点损失可以表示为 L_{p_FL} ,其是更加强调分类不佳的样本的标准二进制交叉熵损失的加权扩展。焦点损失表达式如下:

[0068]
$$L_{p-q} = -\frac{1}{N} \sum_{i=1}^N (1 - p_i)^q \log p_i$$

[0069] 其中正确类别 p_{1i} 的预测概率为

$$[0070] \quad p_{it} = \begin{cases} \hat{Y}_i & , Y_i = 1 \\ 1 - \hat{Y}_i & , Y_i = 0 \end{cases}$$

[0071] 方案3

[0072] γ 为实参数,其被设置为1。在二进制情况下, $Y_i \in \{0, 1\}$ 是地面实况标签,并且 $\hat{Y}_i$ 是第 i 个样本的预测呈递概率。

[0073] 排名目标可进一步用于细胞表面肽呈递数据集以进行训练。例如,N向分类可用于面向排名的训练。N向分类允许阳性样本与阴性样本在数据集中进行竞争。然后可以对样本进行分类,其中每组N个样本都是正样本($N = \text{阴性样本的数量} + 1$)。对于N向分类损失,可以应用交叉熵或焦点损失函数。

[0074] 细胞表面肽呈递数据集可包含至少约5,000个、约10,000个、约15,000个、约20,000个、约25,000个、约30,000个、约35,000个、约40,000个、约45,000个、约50,000个、约60,000个、约70,000个、约80,000个、约90,000个、约100,000个、约150,000个、约200,000个、约250,000个、约300,000个、约350,000个、约400,000个、约450,000个、约500,000个、约550,000个、约600,000个、约650,000个、约700,000个、约750,000个、约800,000个、约850,000个、约900,000个、约950,000个、约1,000,000个、约1,250,000个、约1,500,000个、约1,750,000个、约2,000,000个或更多个肽。大于35,000个样本的训练数据集是优选的。

[0075] 神经网络模型可以在所有训练数据或其一部分上进行训练。例如,可以在约100%的训练数据、约95%、约90%、约85%、约80%、约75%、约70%、约65%、约60%、约55%或更少的训练数据集上训练神经网络模型。可以在MHC I类或MHC II类亲和力测量集中的所有训练数据和细胞表面肽呈递训练数据集中的所有训练数据上训练神经网络模型。例如,可以在约100%、约95%、约90%、约85%、约80%、约75%、约70%、约65%、约60%、约55%或更少的MHC I类或MHC II类亲和力测量数据集和/或约100%、约95%、约90%、约85%、约80%、约75%、约70%、约65%、约60%、约55%或更少的细胞表面肽呈递训练数据集上训练神经网络模型。

[0076] 在一种实现方式中,来自一个或多个训练数据集的训练数据可以进行交叉训练。

例如,MHC I类或MHC II类亲和力测量数据集和细胞表面肽呈递数据集可以进行交叉训练。每个数据集通常包含单个已知目标。例如,MHC I类或MHC II类亲和力测量数据包括肽亲和力,并且细胞表面肽呈递数据集包含可由MHC I类或MHC II类蛋白在细胞表面上呈递的肽。为了对训练数据集进行交叉训练,可以推断每个训练集的目标。例如,可以推断呈递于细胞表面上的肽具有高结合亲和力值,并且可以推断不呈递于细胞表面上的肽具有低结合亲和力。例如,具有高结合亲和力的肽可被推断为在细胞表面上呈递肽,而具有低结合亲和力的肽可被推断为不在细胞表面上呈递肽。

[0077] 在一种实现方式中,可以对来自一个或多个训练数据集的训练数据执行自蒸馏。自蒸馏可以通过提取多个样本的结合亲和力和呈递估计来执行。这些样本可以添加到训练数据集的对应弱标签中。可以使用多等位基因光谱数据进行自蒸馏。可以使用阳性呈现者进行自蒸馏。对于训练数据集中具有未知结合亲和力的阳性呈现者,可以使用已建立的模型来估计结合亲和力。

[0078] 神经网络模型优选地在阳性训练数据和阴性训练数据上进行训练以限制偏差。在不平衡数据集上训练网络可能会通过学习数据主导类的更多表示而使神经网络模型产生偏差,而其他类可能会被忽略。例如,仅在阳性训练数据集上训练的神经网络模型可能会偏向于过度预测肽肿瘤特异性新抗原MHC I类或MHC II类结合亲和力,或过度预测肿瘤特异性新抗原将由MHC I类或MHC II类蛋白在细胞表面上呈递的概率。仅在阴性训练集上训练的神经网络模型可能会偏向于低估肽肿瘤特异性新抗原MHC I类或MHC II类结合亲和力,或低估肿瘤特异性新抗原将由MHC I类或MHC II类蛋白在细胞表面上呈递的概率。

[0079] MHC I类或MHC II类亲和力测量数据集通常包括阳性训练数据和阴性训练数据两者。例如,阳性训练数据可以包括被分类为阳性的结合亲和力预测(例如,<5,000nm的结合亲和力)。例如,阴性训练数据可以包括呈阴性的结合亲和力预测(例如,>5,000nm的结合亲和力)。如果需要的话,则可以通过扩充训练数据集以包含具有低亲和力的随机肽,来将附加的阴性训练数据合并到训练集中。例如,随机肽可以具有约>20,000nm的定性弱亲和力目标。

[0080] 细胞表面肽呈递训练数据集通常包含阳性训练数据(例如,由MHC I类蛋白在细胞表面上呈递的肽)并且不包含阴性训练数据(例如,无法由MHC I类蛋白在细胞表面上呈递的肽)。当训练数据集不包含阴性训练数据时,阳性训练数据集可用于生成概率阴性训练数据集(例如,从阳性训练数据集导出的阴性训练数据集)。可以通过对HLA等位基因的‘阳性’肽进行改组来生成阴性训练数据集。可以通过改变氨基酸长度(例如,使肽更长或更短)来改组肽。或者,可以通过例如氨基酸取代、插入或缺失来修饰肽氨基酸序列。插入包括氨基和/或羧基末端融合以及序列内插入或多个氨基酸残基。缺失的特征在于从肽序列中去除一个或多个氨基酸残基。氨基酸取代通常是单残基取代,但也可以发生在多个位置处。可以组合取代、缺失、插入或它们的任何组合以得到不由MHC I类或MHC II类蛋白在细胞表面上呈递的肽。例如,具有以下氨基酸序列的肽序列:AVGGGERRYIKL可以修饰成:CVGGGEHRYIMNNL。

[0081] 此外,或者与肽改组组合,HLA改组可用于生成阴性训练数据集。被分类为‘阳性’的HLA等位基因(例如,在细胞表面上呈递对应肽的HLA等位基因),可以用不属于阳性等位基因超类型的不同等位基因替换。

[0082] 另外,或者与肽改组和/或HLA改组组合,HRV负采样可用于生成阴性训练数据集。

[0083] 可以进一步过滤训练数据以去除冗余肽。例如,可以去除重复(例如,具有相同氨基酸序列)的肽,使得训练数据集包含独特的肽。本领域技术人员将容易理解如何确定肽的身份(即,确定肽是相同还是不同)。

[0084] 可以使用免疫原性数据集来验证所述经训练的神经网络模型。验证神经网络模型可包括将一个或多个排名度量应用于免疫原性数据集。免疫原性验证数据集中的肽可以基于肽的预测MHC I类或MHC II类结合亲和力以及肽将由MHC I类或MHC II类蛋白在细胞表面上呈递的数值概率进行排名。可以汇总所有等位基因的排名度量。可以通过使用加权等位基因频率来汇总所述排名度量。

[0085] 在实现方式中,可以使用未标记的数据集来训练神经网络模型。例如,细胞表面肽呈递数据集中的肽可能是未标记的。不受理论的束缚,认为未标记的数据集(例如,肽序列)可以提供可更准确地表征所输入的肽序列的数值向量表示。

[0086] III. 模型架构

[0087] 本公开涉及使用神经网络模型来联合预测肿瘤特异性新抗原MHC I类或MHC II类结合亲和力以及对于每种感兴趣的肽,对应肽将由MHC I类或MHC II类蛋白在细胞表面(即,肿瘤细胞的表面)上呈递的数值概率。神经网络模型适合于神经网络模型之前在训练中已遇到过或尚未遇到过的肿瘤特异性新抗原。

[0088] 神经网络模型可以是包括布置在一个或多个层中的一系列节点的单一神经网络。节点可以经由各自具有关联参数的连接部来连接至其他节点。一个特定节点处的值可以被表示为通过由与所述特定节点关联的激活函数所映射的关联参数所加权的连接至所述特定节点的节点的值的总和。本文所述的方法中使用的神经网络模型可以是泛等位基因模型、等位基因特异性模型、超类型特异性模型或它们的组合。

[0089] 在一个具体实施方式中,所述方法包括将肿瘤特异性新抗原的肽序列转换成数值向量。所述肽序列通常表示为字符串,其中每个字符代表氨基酸。肽序列可以转换成数值向量,所述数值向量包含描述肽的氨基酸和氨基酸位置的信息。数值向量可以是二进制分类。例如,具有 k^i 个氨基酸的肽序列 p^i 由20个氨基酸(20-k)的行向量表示,其中对应于所述肽序列的特定位置处的氨基酸的字母表的单个要素的值为1。其余要素的值将为0。作为示例,对于氨基酸字母表A、C、D、E、F、G、H、I、K、L、M、N、P、Q、R、S、T、V、W和Y,4个氨基酸的肽序列AGQY可以由80个要素的行向量表示,并且 $p^i = 10\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 1$ 。

[0090] 肿瘤特异性新抗原的肽序列的长度可为约5个氨基酸至约40个氨基酸。例如,肽序列可为5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个、21个、22个、23个、24个、25个、26个、27个、28个、29个、30个、31个、32个、33个、34个、35个、36个、37个、38个、39个或40个氨基酸。MHC I类分子与短肽结合。MHC I类分子可容纳长度通常为约5个氨基酸至约10个氨基酸的肽。在实施方案中,肿瘤特异性新抗原的肽序列是长度为约5个氨基酸至约10个氨基酸的短肽。MHC II类分子与长度较长的肽结合。MHC II类可容纳长度通常为约13个氨基酸至约25个氨基酸的肽。在实施方案中,肿瘤特异性新抗原的肽序列是长度为约13个至25个氨基酸的长肽。

[0091] 肿瘤特异性新抗原的肽序列可为相同或不同的长度。当肿瘤特异性新抗原的肽序列具有不同长度(例如,一个肽序列是7个氨基酸并且一个肽序列是15个氨基酸)时,可以将填充字符添加到数值向量中,使得直到所述肽的肿瘤特异性新抗原的每个肽达到包括侧翼区的最大肽长度(例如,15个氨基酸)。可以将填充字符添加到侧翼区的C末端或N末端。作为示例,填充字符将编码氨基酸A、C、D、E、F、G、H、I、K、L、M、N、P、Q、R、S、T、V、W和Y。

[0092] 肿瘤特异性新抗原的肽序列可包含在肿瘤特异性新抗原肽侧翼的序列。所述侧翼序列可以直接位于所述肿瘤特异性新抗原肽序列的左侧,直接位于所述肿瘤特异性新抗原肽序列的右侧,或两者。

[0093] 在实施方案中,肿瘤特异性新抗原的肽序列可包含在其源蛋白序列内的肿瘤特异性新抗原肽侧翼的C末端序列中的至少一个C末端序列,或在其源蛋白序列内的肿瘤特异性新抗原肽侧翼的N末端序列中的至少一个N末端序列。优选的是,肿瘤特异性新抗原的肽序列包含在肿瘤特异性新抗原肽侧翼的至少一个C末端氨基酸序列和在肿瘤特异性新抗原肽侧翼的至少一个N末端氨基酸序列。

[0094] 所述侧翼区的长度可为1个、2个、3个、4个、5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个、21个、22个、23个、24个、25个、26个、27个、28个、29个、30个、31个、32个、33个、34个、35个、36个、37个、38个、39个、40个或更多个氨基酸。直接位于肿瘤特异性新抗原肽左侧的侧翼区的长度可为1个、2个、3个、4个、5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个或更多个氨基酸。直接位于肿瘤特异性新抗原肽右侧的侧翼区的长度可为1个、2个、3个、4个、5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个或更多个氨基酸。在实施方案中,肿瘤特异性新抗原的肽序列包含直接位于肿瘤特异性新抗原左侧的长度可为至多约10个氨基酸的侧翼区和/或直接位于肿瘤特异性新抗原右侧的长度可为至多约10个氨基酸的侧翼区。优选地,所述侧翼区包含直接位于肿瘤特异性新抗原的左侧的长度为5个氨基酸的侧翼区,和直接位于肿瘤特异性新抗原的右侧的长度为5个氨基酸的侧翼区。所述侧翼区可以类似地被编码到如上所述的数值向量中。

[0095] 所述方法进一步包括将HLA等位基因伪序列转换成数值向量。HLA等位基因伪序列代表HLA等位基因。HLA等位基因伪序列可为约5个氨基酸至约100个氨基酸。例如,HLA等位基因伪序列可为5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个、21个、22个、23个、24个、25个、26个、27个、28个、29个、30个、31个、32个、33个、34个、35个、36个、37个、38个、39个、40个、41个、42个、43个、44个、45个、46个、47个、48个、49个、50个、51个、52个、53个、54个、55个、56个、57个、58个、59个、60个、61个、62个、63个、64个、65个、66个、67个、68个、69个、70个、71个、72个、73个、74个、75个、76个、77个、78个、79个、80个、81个、82个、83个、84个、85个、86个、87个、88个、89个、90个、91个、92个、93个、94个、95个、96个、97个、98个、99个、或100个氨基酸。HLA等位基因伪序列的长度可为约30至约60个氨基酸。在本文所公开的方法的具体实施方案中,HLA等位基因伪序列的长度为约40至约50个氨基酸。

[0096] 神经网络模型的输入可包含(i)包含肿瘤特异性新抗原的肽序列和所述肽序列的侧翼区的数值向量,以及(ii)包含HLA等位基因伪序列的数值向量。所述神经网络的输入可以任选地包含分段标识符序列。分段标识符序列通知模型每个氨基酸属于哪个区段。接下

来,将包含肿瘤特异性新抗原的肽序列和所述肽序列的侧翼区的数值向量和(ii)包含HLA等位基因伪序列的数值向量编码到一个或多个嵌入层中。所述嵌入层将包含肿瘤特异性新抗原的肽序列和所述肽序列的侧翼区的高维向量以及包含HLA等位基因伪序列的数值向量翻译到低维空间中。所述嵌入层可以被认为是神经网络模型的第一层。然后可以将嵌入层展平以产生肿瘤特异性新抗原的每个肽序列、肽侧翼区和HLA等位基因伪序列的数值向量表示。

[0097] 为了预测肽肿瘤特异性新抗原MHC I类或MHC II类结合亲和力,将肿瘤特异性肽序列和HLA等位基因伪序列拼接。这意味着肿瘤特异性新抗原肽序列和HLA等位基因伪序列以链或串联形式连接在一起。为了预测MHC I类或MHC II类结合亲和力,不需要拼接侧翼区。尽管不是必需的,但在一些情况下可能需要拼接侧翼区。一旦已经拼接肿瘤特异性新抗原肽序列和HLA伪序列,就可以应用一个或多个参数(例如,层和/或函数)。

[0098] 可以应用的示例性层包括但不限于完全连接的致密层、序列层、激活层、归一化层、暂退层、裁剪层、池化层和上池化层、组合层、对象检测层、或生成对抗网络层。

[0099] 示例性完全连接的致密层包括2-D卷积层、3-D卷积层、2D分组卷积层、转置2D-卷积层、转置3-D卷积层、或完全连接的致密层。示例性序列层包括序列输入层、LSTM层、双向LSTM层、GRU层、序列折叠层、序列展开层、展平层或词嵌入层。示例性激活层包括ReLU层、泄漏ReLU层、限幅ReLU层、ELU激活层、双曲正切激活层或PReLU层。示例性的归一化、暂退和裁剪层包括批量归一化层、组归一化层、逐通道局部响应归一化层、暂退层、2-D裁剪层、3-D裁剪层、2-D重置大小层、3-D重置大小层。示例性池化层和上池化层包括平均池化层、3D层、全局平均池化层、3-D全局平均池化层、最大池化层、3-D最大池化层、全局最大池化层或最大上池化层。示例性组合层包括加法层、乘法层、深度拼接层、加权平均层。示例性对象检测层包括ROI输入层、ROI最大池化层、ROI对齐层、锚框层、区域建议层、SSD合并层、空间到深度层、区域建议网络、焦点损失层、区域建议网络、框回归。

[0100] 在实施方案中,可以应用一个或多个完全连接的致密层。所述完全连接层将输入乘以权重矩阵,然后加上偏差向量。例如,可以应用1个、2个、3个、4个、5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个或更多个完全连接的致密层。当预测MHC I类或MHC II类结合亲和力时,优选应用至少3个完全连接的致密层。

[0101] 在实施方案中,可以应用一个或多个激活层(函数)。激活函数可以指派给神经元或神经元的整个层。可以应用的示例性激活函数是ELU激活函数或ReLU层。可以应用上述和/或本领域技术人员已知的其他激活层。激活函数可以将来自节点的求和的加权输入转化为所述节点的激活或输出。例如,可以应用1个、2个、3个、4个、5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个或更多个激活层(函数)。通常,可以应用约1个、2个、3个、4个或5个激活函数。

[0102] 可以应用一个或多个暂退层。暂退层有利于减少过度拟合,并由此提供更好的结果。例如,可以应用1个、2个、3个、4个、5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个或更多个暂退层。通常,可以应用约1个、2个、3个、4个或5个暂退层。

[0103] 在示例性神经网络模型中,可以应用一个或多个完全连接的致密层、一个或多个激活函数和一个或多个暂退层。在优选的神经网络模型中,可以应用一个或多个完全连接

的致密层、激活函数(例如,ELU激活函数)、和一个或多个暂退层。

[0104] 为了使得能够更好地学习序列表示,可以应用一个或多个LSTM层或者一个或多个双向LTSM层。还可以添加转换器。例如,以添加约1个、约2个、约3个、约4个、约5个、约6个、约7个、约8个、约9个、约10个、约11个、约12个、约13个、约14个、约15个或更多个转换器层。转换器可将位置嵌入添加到嵌入的氨基酸序列中,并且可包括一个或多个堆叠的编码器层。例如,转换器可包括一个或多个多头注意层、一个或多个暂退层、一个或多个归一化层、一个或多个前馈层或它们的组合。示例性转换器可包括:(1)多头注意力层、(2)暂退层(0.1比率)、(3)归一化层、(4)前馈层(线性层和ReLU层)、(5)暂退层(0.1比率)、和(6)层归一化)。

[0105] 可以应用回归模型来预测肽肿瘤特异性新抗原MHC I类或II类结合亲和力。特别地,可以使用均方损失函数的变体。均方损失函数可由 L_{BA-MSE} 表示,其中测量与(>)或(<)相关联,仅在不等式被违反时才有助于损失,用于处理数据集中的定量和定性肽-MHC结合亲和力测量两者。可以采用的表达式如方案1和方案2中所示。

[0106] 输出包含代表肽配体MHC I类或MHC II类结合亲和力的数值得分。

[0107] 神经网络模型进一步包括联合预测肿瘤特异性新抗原将由MHC I类或MHC II类蛋白在细胞表面上呈递的概率。为了预测肿瘤特异性新抗原将由MHC I类蛋白在细胞表面上呈递的概率,将肿瘤特异性肽序列、对应的侧翼区和HLA等位基因伪序列拼接成单个数值得分。

[0108] 一旦肿瘤特异性新抗原肽序列、对应的侧翼区和HLA伪序列已被拼接,就可以应用一个或多个参数(例如,层和/或函数)。

[0109] 可以应用一个或多个完全连接的致密层。例如,可以应用1个、2个、3个、4个、5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个或更多个完全连接的致密层。为了预测肿瘤特异性新抗原将由MHC I类或II类蛋白在细胞表面上呈递的概率,优选应用至少3个完全连接的致密层。

[0110] 可以将一个或多个激活函数指派给神经元或神经元的整个层。可以应用的示例性激活函数是ELU激活函数或ReLU层。可以应用上述和/或本领域技术人员已知的其他激活层。激活函数可以将来自节点的求和的加权输入转化为所述节点的激活或输出。例如,可以应用1个、2个、3个、4个、5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个或更多个激活层(函数)。通常,可以应用约1个、2个、3个、4个或5个激活函数。

[0111] 可以应用一个或多个暂退层。暂退层有利于减少过度拟合,并由此提供更好的结果。例如,可以应用1个、2个、3个、4个、5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个或更多个暂退层。通常,可以应用约1个、2个、3个、4个或5个暂退层。

[0112] 可以采用焦点损失二进制分类(参见上面的方案3)来预测肿瘤特异性新抗原将由MHC I类或MHC II类蛋白在细胞表面上呈递的概率。输出是肽将由MHC I类或MHC II类蛋白在细胞表面上呈递的数值概率。

[0113] 可以进一步校准神经网络模型。如果模型未经校准,则神经网络可能会高估或低估概率。因此,校准本文所述的神经网络可以提高预测概率的准确性和置信度。为了校准神经网络,可以应用概率计算。特别地,概率计算可以应用于受试者的HLA等位基因中的一个

或多个HLA等位基因。例如,1个、2个、3个、4个、5个或6个HLA等位基因。概率计算可用于基于模型对每个HLA等位基因的预测来估计受试者的等位基因的总体呈递概率。可以通过校准验证数据集上的神经网络呈递预测来进一步校准神经网络。例如,可以对校准曲线应用低次多项式(low degree polynomial)。可以将多项式系数约束为正以获得单调递增函数。可以使用Lasso线性回归。经校准的呈递预测(例如,肿瘤特异性新抗原MHC I类结合亲和力和肿瘤特异性新抗原将由MHC I类蛋白在细胞表面上呈递的概率)可以用作免疫原性的指标。

[0114] 可以使用本文所述的免疫原性数据来评估神经网络模型的性能。可以使用一个或多个排名度量来评估神经网络预测。示例性排名度量包括但不限于排名前k的项目、Precision@K、 DCG_k 、倒数排名和阳性预测值度量。可以使用一种或多种排名度量基于预测的细胞表面肽呈递概率和/或预测的结合亲和力得分对免疫原性数据集中每个等位基因的所有对应肽进行排名。然后可以使用加权等位基因频率来汇总排名度量。

[0115] 在一种实现方式中,可以执行自监督预训练。示例性的训练模型是掩码语言建模和下一肽预测。

[0116] IV. 方法的计算机实现

[0117] 编程或以其他方式配置的计算机系统可用于实现本文所公开的方法。计算机系统可包括单个计算装置或使用一个或多个计算网络互连的多个计算装置。计算机系统可以使用计算机能力执行本文所述的神经网络模型。

[0118] 计算机系统可包括中央处理单元,所述中央处理单元可以是单核或多核处理器,或者用于并行处理的多个处理器。所述系统可包括存储器(例如,随机存取存储器、只读存储器、闪存)、电子存储单元(例如,云平台)、用于与一个或多个系统通信的通信接口、以及其他外围装置,诸如数据存储装置、其他存储器和显示适配器。存储器、存储装置、接口和外围装置可以通过通信总线与CPU通信。这些部件中的任何一个或全部部件可以经由共享的内部网络或外部网络进行通信,并且集合系统可以经由网络与一个或多个用户装置进行通信。所述网络可以是互联网、外联网、或者与互联网通信的互联网/外联网。所述网络可包括一个或多个计算机服务器,所述一个或多个计算机服务器可以使得能够实现分布式计算,诸如云计算。所述计算机系统可以与处理系统通信。所述处理系统可被配置为实现本文所公开的方法。

[0119] 计算装置的各种示例可包括但不限于桌上型计算机、膝上型计算机和移动电话、平板计算机、个人计算机、可穿戴计算机、服务器、个人数字助理(PDA)、混合PDA/移动电话、移动电话、电子书阅读器、机顶盒、语音命令装置、相机、数字媒体播放器等。在一些实施方案中,计算机装置可以具有一个或多个用户接口、命令行接口(CLI)、应用程序编程接口(API)、和/或用于提交训练请求、部署请求和/或执行请求的其他编程接口。在一些实施方案中,计算机装置可以执行与神经网络模型交互的独立应用程序。

[0120] 在一些实施方案中,所述网络包括任何有线网络、无线网络或它们的组合。例如,所述网络可以是域网、局域网、广域网、无线广播网络(例如,用于无线电或电视)、电缆网络、卫星网络、蜂窝电话网络或它们的组合。作为另外的示例,所述网络可以是链接网络的公共可访问网络,可能由各种不同的各方(诸如互联网)操作。在一些实施方案中,网络可以是私有或半私有网络,诸如公司或大学内联网。所述网络可包括一个或多个无线网络,诸如全球移动通信系统(GSM)网络、码分多址(CDMA)网络、长期演进(LTE)网络或任何其他类型

的无线网络。网络可以使用用于经由互联网或任何其他前述类型的网络进行通信的协议和部件。例如,网络所使用的协议可包括HTTP、HTTP安全(HTTPS)、消息队列遥测传输(MQTT)、受限应用程序协议(CoAP)等。用于经由互联网或任何其他前述类型的通信网络进行通信的协议和部件对于本领域技术人员来说是众所周知的,并且因此在本文中不再更详细描述。

[0121] 图9示出了根据一些实施方案的示例性提供商网络(或“服务提供商系统”)环境。提供商网络900可以经由一个或多个虚拟化服务910向客户提供资源虚拟化,所述虚拟化服务允许客户购买、租用或以其他方式获得虚拟化资源的实例912,包括但不限于在一个或多个数据中心中的一个或多个提供商网络内的装置上实现计算和存储资源。本地互联网协议(IP)地址916可以与资源实例912相关联;本地IP地址是提供商网络900上的资源实例912的内部网络地址。在一些实施方案中,提供商网络900还可以提供客户可以从提供商900获得的公共IP地址914和/或公共IP地址范围(例如,互联网协议版本4(IPv4)或互联网协议版本6(IPv6)地址)。

[0122] 按照惯例,提供商网络900可经由虚拟化服务910允许服务提供商的客户(例如,操作包括一个或多个客户装置952的一个或多个客户网络950A-950C的客户)动态地将指派或分配给客户的至少一些公共IP地址914与指派给客户的特定资源实例912相关联。提供商网络900还可以允许客户将先前映射到分配给客户的一个虚拟化计算资源实例912的公共IP地址914重新映射到也分配给客户的另一虚拟化计算资源实例912。使用由服务提供商提供的虚拟化计算资源实例912和公共IP地址914,服务提供商的客户(诸如客户网络950A-950C的运营商)可以例如实现客户特定的应用程序并在中间网络940上(诸如互联网)呈现客户的应用程序。然后,中间网络940上的其他网络实体920可以生成到由客户网络950A-950C发布的目的地公共IP地址914的通信量;所述通信量被路由到服务提供商数据中心,并且在数据中心处经由网络底层路由到当前映射到目的地公共IP地址914的虚拟化计算资源实例912的本地IP地址916。类似地,来自虚拟化计算资源实例912的响应通信量可以经由网络底层路由回到中间网络940上以到达源实体920。

[0123] 如本文所用,本地IP地址是指例如提供商网络中的资源实例的内部或“私有”网络地址。本地IP地址可以位于互联网工程任务组(IETF)请求注解(RFC)1918保留的地址块内和/或具有由IETF RFC 4193指定的地址格式内,并且可以在提供商网络内可变。源自提供商网络外部的网络通信量不会直接路由到本地IP地址;相反,所述通信量使用映射到资源实例的本地IP地址的公共IP地址。提供商网络可包括提供网络地址转换(NAT)或类似功能性以执行从公共IP地址至本地IP地址的映射并且反之亦然的网络装置或器具。

[0124] 公共IP地址是由服务提供商或客户指派给资源实例的互联网可变网络地址。路由至公共IP地址的通信量经转换(例如经由1:1NAT),并转发至资源实例的相应本地IP地址。

[0125] 一些公共IP地址可以由提供商网络基础设施指派给特定资源实例;这些公共IP地址可称为标准公共IP地址,或者简称为标准IP地址。在一些实施方案中,标准IP地址至资源实例的本地IP地址的映射是所有资源实例类型的默认启动配置。

[0126] 至少一些公共IP地址可以被分配给提供商网络900的客户或者由所述客户获得;然后,客户可以将其所分配的公共IP地址指派至分配给所述客户的特定资源实例。这些公共IP地址可被称为客户公共IP地址,或简称为客户IP地址。代替如在标准IP地址的情况下由提供商网络900指派至资源实例,客户IP地址可以由客户例如经由由服务提供商提供的

API指派至资源实例。与标准IP地址不同,客户IP地址被分配给客户帐户,并且可以由相应客户根据需要或按期望重新映射到其他资源实例。客户IP地址与客户帐户相关联,而不是与特定的资源实例相关联,并且客户控制该IP地址,直到客户选择释放IP地址。与传统的静态IP地址不同,客户IP地址允许客户通过将客户的公共IP地址重新映射到与客户帐户关联的任何资源实例来掩蔽资源实例或可用区故障。例如,客户IP地址使客户能够通过将客户IP地址重新映射至替换资源实例来围绕客户的资源实例或软件问题进行工程化。

[0127] 图10是根据一些实施方案的向客户提供存储服务 and 硬件虚拟化服务的示例性提供商网络的框图。硬件虚拟化服务1020向客户提供多个计算资源1024(例如,VM)。计算资源1024可以例如被出租或租赁给提供商网络1000的客户(例如,给实现客户网络1050的客户)。每个计算资源1024可具有一个或多个本地IP地址。提供商网络1000可被配置为将数据包从计算资源1024的本地IP地址路由至公共互联网目的地,以及将数据包从公共互联网源路由到计算资源1024的本地IP地址。

[0128] 提供商网络1000可以向例如经由本地网络1056耦接至中间网络1040的客户网络1050提供经由耦接至中间网络1040和提供商网络1000的硬件虚拟化服务1020实现虚拟计算系统1092的能力。在一些实施方案中,硬件虚拟化服务1020可以提供一个或多个API 1002,例如网络服务接口,经由所述API,客户网络1050可以例如经由控制台1094(例如,基于网络的应用程序、独立应用程序、移动应用程序等)访问由硬件虚拟化服务1020提供的功能性。在一些实施方案中,在提供商网络1000处,客户网络1050处的每个虚拟计算系统1092可以对应于租赁、出租或以其他方式提供给客户网络1050的计算资源1024。

[0129] 从虚拟计算系统1092和/或另一客户装置1090的实例(例如,经由控制台1094),客户可以例如经由一个或多个API 1002来访问存储服务1010的功能,以访问来自提供商网络1000提供的虚拟数据存储装置1016(例如,文件夹或“桶”、虚拟卷、数据库等)的存储资源1018A-1018N的数据并将数据存储至所述存储资源。在一些实施方案中,可以在客户网络1050处提供虚拟化数据存储网关(未示出),所述虚拟化数据存储网关可以本地高速缓存至少一些数据,例如频繁访问的数据或关键数据,并且可以经由一个或多个通信通道与存储服务1010通信以从本地高速缓存上传新的或修改的数据,以便维护数据的主存储装置(虚拟化数据存储装置1016)。在一些实施方案中,经由虚拟计算系统1092和/或在另一客户装置1090上,用户可以经由充当存储虚拟化服务的存储服务1010安装和访问虚拟数据存储装置1016卷,并且这些卷可以向用户显现作为本地(虚拟)存储装置1098。

[0130] 虽然在图10中未示出,但是还可以经由API 1002从提供商网络1000内的资源实例访问虚拟化服务。例如,客户、设备服务提供商或其他实体可以经由API 1002从提供商网络1000上的相应虚拟网络内访问虚拟化服务,以请求在所述虚拟网络内或在另一虚拟网络内分配一个或多个资源实例。

[0131] 在一些实施方案中,实现本文所述的技术的一部分或全部的系统可以包括通用计算机系统,所述通用计算机系统包括或被配置为访问一个或多个计算机可访问介质,诸如图11所示的计算机系统1100。在所示实施方案中,计算机系统1100包括经由输入/输出(I/O)接口1130耦接至系统存储器1120的一个或多个处理器1110。计算机系统1100进一步包括耦接至I/O接口1130的网络接口1140。虽然图11将计算机系统1100示出为单个计算装置,但是在各种实施方案中,计算机系统1100可包括一个计算装置或被配置为一起作为单个计算

机系统1100工作的任何数量的计算装置。

[0132] 在各种实施方案中,计算机系统1100可以是包括一个处理器1110的单处理器系统,或者包括若干处理器1110(例如,两个、四个、八个或另一合适数量)的多处理器系统。处理器1110可以是能够执行指令的任何合适的处理器。例如,在各种实施方案中,处理器1110可以是实现多种指令集架构(ISA)中的任何一种指令集架构的通用或嵌入式处理器,诸如x86、ARM、PowerPC、SPARC或MIPS ISA,或任何其他合适的ISA。在多处理器系统中,处理器1110中的各个处理器可以共同但不一定实现相同的ISA。

[0133] 系统存储器1120可以存储可由处理器1110访问的指令和数据。在各种实施方案中,系统存储器1120可以使用任何合适的存储器技术,诸如随机存取存储器(RAM)、静态RAM(SRAM)、同步动态RAM(SDRAM)、非易失性/闪存型存储器或任何其他类型的存储器来实现。在所示实施方案中,实现一种或多种期望功能的程序指令和数据,诸如上述那些方法、技术和数据,被示出为存储在系统存储器1120内。

[0134] 在一个实施方案中,I/O接口1130可被配置为协调处理器1110、系统存储器1120和装置中的任何外围装置(包括网络接口1140或其他外围接口)之间的I/O通信量。在一些实施方案中,I/O接口1130可以执行任何必要的协议、定时或其他数据转换,以将来自一个部件(例如,系统存储器1120)的数据信号转换成适合于由另一部件(例如,处理器1110)使用的格式。在一些实施方案中,I/O接口1130可以包括对通过各种类型的外围总线(诸如例如外围部件互连(PCI)总线标准或通用串行总线(USB)标准的变体)附接的装置的支持。在一些实施方案中,I/O接口1130的功能可以被分成两个或更多个单独的部件,例如北桥和南桥。此外,在一些实施方案中,I/O接口1130的一些或全部功能,诸如至系统存储器1120的接口,可以直接并入处理器1110中。

[0135] 网络接口1140可被配置为允许在计算机系统1100和附接至一个或多个网络1150的其他装置1160(诸如如图1所示的其他计算机系统或装置)之间交换数据。在各种实施方案中,网络接口1140可以支持经由任何合适的有线或无线通用数据网络(诸如各种类型的以太网)进行通信。另外,网络接口1140可以支持经由诸如模拟语音网络或数字光纤通信网络的电信/电话网络、诸如光纤通道SAN的存储区域网络(SAN)、或者经由I/O任何其他合适类型的网络和/或协议进行通信。

[0136] 在一些实施方案中,计算机系统1100包括使用I/O接口1130(例如,实现高速外围部件互连(PCI-E)标准版本或另一互连(诸如QuickPath互连(QPI)或UltraPath互连(UPI))的总线)连接的一个或多个卸载卡1170(包括一个或多个处理器1175,并且可能包括一个或多个网络接口1140)。例如,在一些实施方案中,计算机系统1100可以充当托管计算实例的主控端电子装置(例如,作为硬件虚拟化服务的一部分操作),并且一个或多个卸载卡1170执行可以管理在可控端电子装置上执行的计算实例的虚拟化管理器。作为示例,在一些实施方案中,卸载卡1170可以执行计算实例管理操作,诸如暂停和/或取消暂停计算实例、启动和/或终止计算实例、执行转储/复制操作等。在一些实施方案中,这些管理操作可以由卸载卡1170与由计算机系统1100的其他处理器1110A-1110N执行的管理程序(例如,根据来自管理程序的请求)协同执行。然而,在一些实施方案中,由卸载卡1170实现的虚拟化管理器可以适应来自其他实体(例如,来自计算实例本身)的请求,并且可以与任何单独的管理程序协同(或不服务任何单独的管理程序)。

[0137] 在一些实施方案中,系统存储器1120可以是配置为存储如上所述的程序指令和数据的计算机可访问介质的一个实施方案。然而,在其他实施方案中,可以在不同类型的计算机可访问介质上接收、发送或存储程序指令和/或数据。一般而言,计算机可访问介质可包括非暂时性存储介质(storage media/memory media),诸如磁或光介质,例如经由I/O接口1130耦接至计算机系统1100的盘或DVD/CD。非暂时性计算机可访问存储介质还可包括可被包括在计算机系统1100的一些实施方案中作为系统存储器1120或另一类型的存储器的任何易失性或非易失性介质,诸如RAM(例如,SDRAM、双倍数据速率(DDR) SDRAM、SRAM等)、只读存储器(ROM)等。此外,计算机可访问介质可包括经由通信介质(诸如网络和/或无线链路)传送的传输介质或信号(诸如电信号、电磁信号或数字信号),诸如可以经由网络接口1140来实现。

[0138] 本文所讨论或建议的各种实施方案可以在各种操作环境中实现,所述操作环境在一些情况下可包括可用于操作多种应用程序中的任何一种应用程序的一个或多个用户计算机、计算装置或处理装置。用户或客户端装置可包括多种通用个人计算机中的任何一种通用个人计算机,诸如运行标准操作系统的桌上型计算机或膝上型计算机,以及运行移动软件并能够支持多种网络连接和消息传输协议的蜂窝、无线和手持装置。此类系统还可包括运行多种市售操作系统中的任何一种市售操作系统的多个工作站,以及用于诸如开发和数据库管理的目的的其他已知应用程序。这些装置还可包括其他电子装置,诸如虚拟终端、瘦客户端、游戏系统和/或能够经由网络通信的其他装置。

[0139] 大多数实施方案利用本领域技术人员熟悉的至少一个网络来支持使用多种广泛可用的协议中的任何一种协议进行通信,所述协议为诸如传输控制协议/互联网协议(TCP/IP)、文件传输协议(FTP)、通用即插即用(UPnP)、网络文件系统(NFS)、通用互联网文件系统(CIFS)、可扩展消息和表示协议(XMPP)、AppleTalk等。所述网络可包括例如局域网(LAN)、广域网(WAN)、虚拟专用网(VPN)、互联网、内联网、外联网、公共交换电话网(PSTN)、红外网络、无线网络以及它们的任意组合。

[0140] 在利用网络服务器的实施方案中,所述网络服务器可以运行多种服务器或中间层应用程序中的任何一者,包括HTTP服务器、文件传输协议(FTP)服务器、通用网关接口(CGI)服务器、数据服务器、Java服务器、业务应用程序服务器等。服务器还能够响应于来自用户装置的请求来执行程序或脚本,诸如通过执行一个或多个网络应用程序,所述网络应用程序可以被实现为以任何编程语言(诸如Java®、C、C#或C++)或任何脚本语言(诸如Perl、Python、PHP或TCL)以及它们的组合编写的一个或多个脚本或程序。服务器还可包括数据库服务器,包括但不限于可从Oracle(R)、Microsoft(R)、Sybase(R)、IBM(R)等商购获得的那些数据库服务器。数据库服务器可以是关系型或非关系型(例如,“NoSQL”)、分布式或非分布式等。

[0141] 本文所公开的环境可包括如上所述的各种数据存储装置和其他存储器和存储介质。它们可以驻留在多个位置,诸如在一台或多台计算机本地(和/或驻留在在一台或多台计算机中)的存储介质上,或者跨网络与任何或所有计算机远程的存储介质上。在一组特定的实施方案中,信息可以驻留在本领域技术人员熟悉的存储区域网络(SAN)中。类似地,用于执行归因于计算机、服务器或其他网络装置的功能的任何必要的文件可以适当地本地和/或远程存储。当系统包括计算机化的装置时,每个此类装置可包括可以经由总线电耦接

的硬件元件,所述元件包括例如至少一个中央处理单元(CPU)、至少一个输入装置(例如,鼠标、键盘、控制器、触摸屏或小键盘)和/或至少一个输出装置(例如,显示装置、打印机或扬声器)。此类系统还可以包括一个或多个存储装置,诸如磁盘驱动器、光学存储装置和固态存储装置(诸如随机存取存储器(RAM)或只读存储器(ROM))以及可移动介质装置、存储卡、闪存卡等。

[0142] 此类装置还可包括如上所述的计算机可读存储介质读取器、通信装置(例如,调制解调器、网卡(无线或有线)、红外通信装置等)和工作存储器。计算机可读存储介质读取器可以与计算机可读存储介质连接或被配置为接收计算机可读存储介质,该计算机可读存储介质代表远程、本地、固定和/或可移除存储装置以及用于暂时和/或更永久地包含、存储、传输和检索计算机可读信息的存储介质。系统和各种装置通常还包括位于至少一个工作存储器装置内的多个软件应用程序、模块、服务或其他元件,包括操作系统和应用程序,诸如客户端应用程序或网络浏览器。应当理解的是,替代实施方案可以具有与上述实施方案不同的多种变化。例如,还可以使用定制硬件,和/或可以以硬件、软件(包括便携式软件,诸如小应用程序)或两者来实现特定元件。进一步,可以采用至其他计算装置(例如网络输入/输出装置)的连接。

[0143] 用于包含代码或代码的部分的存储介质和计算机可读介质可以包括本领域已知或使用的任何适当的介质,包括存储介质和通信介质,诸如但不限于以任何方法或技术实现的用于存储和/或传输信息(诸如计算机可读指令、数据结构、程序模块或其他数据)的易失性和非易失性、可移除和不可移除介质,包括RAM、ROM、电可擦除可编程只读存储器(EEPROM)、闪存或其他存储技术、光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储装置、磁带盒、磁带、磁盘存储装置或其他磁性存储装置,或任何其他可用于存储所需信息并可由系统装置访问的介质。基于本文所提供的公开内容和教导,本领域普通技术人员将理解实现各种实施方案的其他方式和/或方法。

[0144] 在前面的描述中,描述了各种实施方案。出于解释的目的,阐述了具体配置和细节以便提供对实施方案的透彻理解。然而,对于本领域技术人员来说将显而易见的是,可以在没有具体细节的情况下实践所述实施方案。此外,可以省略或简化众所周知的特征,以免模糊所描述的实施方案。

[0145] 带括号的文本和具有虚线边框(例如,大破折线、小破折线、点划线和点线)的块在本文中用于说明向一些实施方案添加附加特征的任选操作。然而,这种表示法不应被理解为意味着这些是唯一的选项或任选的操作,和/或具有实线边框的块在某些实施方案中不是任选的。

[0146] V. 免疫原性组合物

[0147] 本发明进一步涉及个性化的(即,受试者特异性的)免疫原性组合物(例如,癌症疫苗),所述个性化的免疫原性组合物包含使用本文所述的方法选择的一种或多种肿瘤特异性抗原。此类免疫原性组合物可以根据本领域的标准规程配制。所述免疫原性组合物能够引发特异性免疫应答。

[0148] 所述免疫原性组合物可以配制成使得所述肿瘤特异性新抗原的选择和数量是根据受试者的特定癌症定制的。例如,肿瘤特异性新抗原的选择可以取决于癌症的具体类型、癌症的阶段、受试者的免疫状态和受试者的MHC类型。

[0149] 所述免疫原性组合物可包含至少1种、2种、3种、4种、5种、6种、7种、8种、9种、10种、11种、12种、13种、14种、15种、16种、17种、18种、19种、20种、21种、22种、23种、24种、25种、26种、27种、28种、29种、30种、31种、32种、33种、34种、35种、36种、37种、38种、39种、40种、41种、42种、43种、44种、45种、46种、47种、48种、49种、50种或更多种肿瘤特异性新抗原。所述免疫原性组合物可含有约10-20种肿瘤特异性新抗原、约10-30种肿瘤特异性新抗原、约10-40种肿瘤特异性新抗原、约10-50种肿瘤特异性新抗原、约10-60种肿瘤特异性新抗原、约10-70种肿瘤特异性新抗原、约10-80种肿瘤特异性新抗原、约10-90种肿瘤特异性新抗原、或约10-100种肿瘤特异性新抗原。优选地,所述免疫原性组合物包含至少约10种肿瘤特异性新抗原或至少约20种肿瘤特异性新抗原。

[0150] 所述免疫原性组合物可进一步包含天然或合成抗原。所述天然或合成抗原可以增强免疫反应。示例性天然或合成抗原包括但不限于泛DR表位(PADRE)和破伤风毒素抗原。

[0151] 所述免疫原性组合物可以是任何形式,例如合成的长肽、RNA、DNA、细胞、树突细胞、核苷酸序列、多肽序列、质粒或载体。

[0152] 所述肿瘤特异性新抗原也可以包含在基于病毒载体的疫苗平台中,所述基于病毒载体的疫苗平台为诸如牛痘、禽痘、自我复制甲病毒、马拉巴病毒、腺病毒(参见例如,Tatsis等人,Molecular Therapy,10:616-629(2004))、或慢病毒,包括但不限于第二代、第三代或混合的第二/第三代慢病毒和设计用于靶向特定细胞类型或受体的任何代的重组慢病毒(参见例如Hu等人,Immunol Rev.,239(1):45-61(2011),Sakma等人,Biochem J.,443(3):603-18(2012))。取决于上述基于病毒载体的疫苗平台的包装能力,这种方法可以递送编码一种或多种肿瘤特异性新抗原肽的一种或多种核苷酸序列。序列侧翼可以是非突变序列,可以被接头分开,或者可以在其前面有一个或多个靶向亚细胞区室的序列(参见例如,Gros等人,Nat Med.,22(4):433-8(2016),Stronen等人,Science.,352(6291):1337-1341(2016),Lu等人,Clin Cancer Res.,20(13):3401-3410(2014))。在引入宿主中后,感染细胞表达一种或多种肿瘤特异性新抗原,并由此引发针对一种或多种肿瘤特异性新抗原的宿主免疫(例如,CD8⁺或CD4⁺)应答。在免疫方案中有用的牛痘病毒载体和方法描述于例如美国专利号4,722,848中。另一种载体是BCG(卡介苗)。BCG载体描述于Stover等人(Nature 351:456-460(1991))中。还可以使用对于本领域技术人员来说根据本文的描述将是显而易见的可用于新抗原的治疗性施用或免疫的多种其他疫苗载体。

[0153] 根据特定受试者的个人需要,所述免疫原性组合物可含有个体化的组分。

[0154] 本文所述的免疫原性组合物可进一步包含佐剂。佐剂是这样的任何物质,所述物质与免疫原性组合物的掺和物增加或以其他方式增强和/或提升针对肿瘤特异性新抗原的免疫应答,但是当单独施用时不产生针对肿瘤特异性新抗原的免疫应答。所述佐剂优选地产生针对新抗原的免疫应答并且不产生过敏或其他不良反应。本文预期的是所述免疫原性组合物可以在施用免疫原性组合物之前、一起、同时或之后施用。

[0155] 佐剂可以通过几种机制来增强免疫应答,所述机制包括例如淋巴细胞募集、B和/或T细胞刺激以及巨噬细胞的刺激。当本发明的免疫原性组合物包含佐剂或与一种或多种佐剂一起施用时,可以使用的佐剂包括但不限于矿物盐佐剂或矿物盐凝胶佐剂、微粒佐剂、微米微粒佐剂、粘膜佐剂和免疫刺激佐剂。佐剂的示例包括但不限于铝盐(alum)(诸如氢氧化铝、磷酸铝盐和硫酸铝盐)、3脱氧酰化单磷酰脂质A(MPL)(参见,GB 2220211)、MF59

(Novartis)、AS03 (Glaxo SmithKline)、AS04 (Glaxo SmithKline)、聚山梨醇酯80 (Tween 80; ICL Americas, Inc.)、咪唑并吡啶化合物(参见作为国际公开号W02007/109812公布的国际申请号PCT/US2007/064857)、咪唑并喹啉化合物(参见作为国际公开号W02007/109813公布的国际申请号PCT/US2007/064858)和皂苷,诸如QS21(参见Kensil等人,在Vaccine Design: The Subunit and Adjuvant Approach(编辑Powell和Newman, Plenum Press, NY, 1995)中;美国专利号5,057,540)。在一些实施方案中,佐剂是弗氏佐剂(完全或不完全)。其他合适的佐剂是水包油乳剂(诸如角鲨烯或花生油),任选地与免疫刺激剂诸如单磷酸脂质A的组合(参见Stoute等人, N. Engl. J. Med. 336, 86-91 (1997))。

[0156] CpG免疫刺激性寡核苷酸也已被报道为增强佐剂在疫苗环境中的效应。也可以使用其他TLR结合分子,诸如结合RNA的TLR 7、TLR 8和/或TLR 9。

[0157] 有用佐剂的其他示例包括但不限于经化学修饰的CpG(例如CpR, Idera)、Poly (I: C)(例如polyi:CI2U)、poly ICLC、非CpG细菌DNA或RNA以及免疫活性小分子和抗体,诸如环磷酸胺、舒尼替尼、贝伐单抗、Celebrex(塞来昔布)、NCX-4016、西地那非、他达拉非、伐地那非、索拉非尼、XL-999、CP-547632、帕唑帕尼(pazopamb)、ZD2171、AZD2171、易普利姆玛单抗、曲美木单抗和SC58175,它们可以起到治疗作用和/或充当佐剂。在实施方案中,Poly ICLC是优选的佐剂。

[0158] 所述免疫原性组合物可包含单独或与药学上可接受的载体一起的本文所述的一种或多种肿瘤特异性新抗原。可以使用一种或多种肿瘤特异性新抗原的悬浮液或分散体,尤其是等渗水悬浮液、分散体或两亲性溶剂。免疫原性组合物可以是灭菌的和/或可包含辅料,例如防腐剂、稳定剂、润湿剂和/或乳化剂、增溶剂、用于调节渗透压的盐和/或缓冲剂,并且以本身已知的方式制备,例如借助于常规的分散和悬浮方法。在某些实施方案中,此类分散体或悬浮液可包含粘度调节剂。将悬浮液或分散体保持在约2°C至8°C的温度下,或者优先地为了长期储存可以冷冻,然后在使用前不久解冻。对于注射,疫苗或免疫原性制备物可以配制在水溶液中,优选在生理相容的缓冲液,诸如汉克氏溶液(Hanks's solution)、林格氏溶液或生理盐水缓冲液中。所述溶液可含有配制剂,诸如悬浮剂、稳定剂和/或分散剂。

[0159] 在某些实施方案中,本文所述的组合物另外包含防腐剂,例如汞衍生物硫柳汞。在具体实施方案中,本文所述的药物组合物包含0.001%至0.01%的硫柳汞。在其他实施方案中,本文所述的药物组合物不包含防腐剂。

[0160] 辅料可以独立于佐剂而存在。所述辅料的功能可以是例如增加免疫原性组合物的分子量、增强活性或免疫原性、赋予稳定性、增强生物活性或延长血清半衰期。辅料也可用于帮助将一种或多种肿瘤特异性新抗原呈递给T细胞(例如,CD4⁺或CD8⁺T细胞)。所述辅料可以是载体蛋白,诸如但不限于匙孔血蓝蛋白、血清蛋白(诸如转铁蛋白、牛血清白蛋白、人血清白蛋白)、甲状腺球蛋白或卵清蛋白、免疫球蛋白或激素(诸如胰岛素)或棕榈酸。对于人类的免疫来说,载体通常是人类可接受且安全的生理学上可接受的载体。或者,载体可以是葡聚糖,例如琼脂糖凝胶。

[0161] 细胞毒性T细胞识别呈与MHC分子结合的肽形式的抗原,而不是完整的外来抗原本身。MHC分子位于抗原呈递细胞的细胞表面处。因此,如果存在肽抗原、MHC分子和抗原呈递细胞(APC)的三聚体复合物,则细胞毒性T细胞的激活是可能的。如果不仅使用一种或多种肿瘤特异性抗原来激活细胞毒性T细胞,而且还添加带有相应MHC分子的附加APC,则可能增

强免疫应答。因此,在一些实施方案中,免疫原性组合物另外含有至少一种APC。

[0162] 免疫原性组合物可包含可接受的载体(例如,水性载体)。可以使用多种水性载体,例如水、缓冲水、0.9%盐水、0.3%甘氨酸、透明质酸等。这些组合物可以通过常规的、众所周知的灭菌技术来灭菌,或者可以进行无菌过滤。所得水溶液可经包装以按原样使用,或冻干,所述冻干制备物在施用前与无菌溶液组合。所述组合物可视需要含有药学上可接受的助剂物质以接近生理条件,诸如pH调节和缓冲剂、张力调节剂、润湿剂等,例如乙酸钠、乳酸钠、氯化钠、氯化钾、氯化钙、脱水山梨糖醇单月桂酸酯、三乙醇胺油酸酯等。

[0163] 新抗原也可以经由脂质体施用,所述脂质体将所述新抗原靶向特定的细胞组织,诸如淋巴组织。脂质体还可用于延长半衰期。脂质体包括乳液、泡沫、胶束、不溶性单层、液晶、磷脂分散体、层状层等。在这些制备物中,待递送的新抗原作为脂质体的一部分,单独或与结合例如淋巴细胞中普遍存在的受体的分子(诸如结合CD45抗原的单克隆抗体)或与其他治疗性或免疫原性组合物组合掺入。因此,填充有所需新抗原的脂质体可以被引导至淋巴样细胞的位点,然后脂质体在所述位点处递送所选定的免疫原性组合物。脂质体可以由标准的形成囊泡的脂质形成,所述囊泡形成脂质通常包括中性和带负电的磷脂和甾醇,诸如胆固醇。脂质的选择通常通过考虑例如脂质体大小、酸不稳定性和脂质体在血流中的稳定性来指导。多种方法可用于制备脂质体,如在例如Szoka等人, *An.Rev.Biophys.Bioeng.* 9;467 (1980);美国专利号4,235,871、4,501,728、4,501,728、4,837,028、和5,019,369中所述。

[0164] 为了靶向免疫细胞,待掺入脂质体中的配体可包括例如特异于所需免疫系统细胞的细胞表面决定簇的抗体或其片段。脂质体悬浮液可以以尤其根据施用方式、所递送的肽和所治疗的疾病的阶段而变化的剂量静脉内地、局部性地、局部地等施用。

[0165] 用于靶向免疫细胞的替代方法,可以将免疫原性组合物的组分,诸如抗原(即,肿瘤特异性新抗原)、配体或佐剂(例如,TLR)掺入聚(乳酸-共聚-乙醇酸)微球中。聚(乳酸-共聚-乙醇酸)微球可以捕集免疫原性组合物的组分作为核内体递送装置。

[0166] 出于治疗或免疫目的,也可以将编码本文所述的肿瘤特异性新抗原的核酸施用于受试者。许多方法可方便地用于将核酸递送至受试者。例如,核酸可以作为“裸DNA”直接递送。这种方法描述于例如Wolff等人, *Science* 247:1465-1468 (1990),以及美国专利号5,580,859和5,589,466中。还可以使用如例如在美国专利号5,204,253中所述的弹道递送来施用核酸。可以施用仅由DNA构成的粒子。或者,DNA可以粘附至粒子,诸如金粒子。用于递送核酸序列的方法可包括在使用或不使用电穿孔的情况下的病毒载体、mRNA载体和DNA载体。核酸还可以与阳离子化合物(诸如阳离子脂质)复合递送。

[0167] 本文所提供的免疫原性组合物可以通过包括但不限于口服、皮内、瘤内、肌内、腹膜内、静脉内、局部、皮下、经皮、鼻内和吸入途径以及经由划痕法(例如使用分叉针划穿皮肤表层)施用于受试者。可以在肿瘤部位施用免疫原性组合物以诱导对肿瘤的局部免疫应答。

[0168] 一种或多种肿瘤特异性新抗原的剂量可以取决于组合物的类型以及受试者的年龄、体重、体表面积、个体状况、个体药代动力学数据和施用模式。

[0169] 本文还公开了一种制造免疫原性组合物的方法,所述免疫原性组合物包含通过执行本文所公开的方法的步骤而选择的一种或多种肿瘤特异性新抗原。可以使用本领域已知

的方法来制造如本文所述的免疫原性组合物。例如,本文所公开的产生肿瘤特异性新抗原或载体(例如,包含至少一个编码一种或多种肿瘤特异性新抗原的序列的载体)的方法可包括在适合表达新抗原或载体的条件下培养宿主细胞,其中所述宿主细胞包含至少一种编码所述新抗原或载体的多核苷酸;以及纯化所述新抗原或载体。标准纯化方法包括色谱技术、电泳、免疫学、沉淀、透析、过滤、浓缩和色谱聚焦技术。

[0170] 宿主细胞可包括中国仓鼠卵巢 (CHO) 细胞、NS0细胞、酵母或HEK293细胞。宿主细胞可以用包含至少一种编码本文所公开的一种或多种肿瘤特异性新抗原或载体的核酸序列的一种或多种多核苷酸转化。在某些实施方案中,分离的多核苷酸可以是cDNA。

[0171] 5. 实施例

[0172] 实施例1.1: 训练数据

[0173] 训练模型以预测肽-MHC结合和MHC I类上内源性肽呈递的概率。这些被视为CD8+T细胞免疫原性的指标。使用来自MHCflurry(“curated_training_data.no_mass_spec.csv” 1)的精选肽-MHC结合亲和力数据,所述数据包含来自IEDB[1]和Kim等人[2]的数据。在这个精选数据集上执行的唯一处理步骤是添加肽源蛋白,以提取侧翼区并用于我们的负采样方法。在这个处理步骤之后,用于训练的最终数据集被称为“curated_training_data.no_mass_spec.multiple_context.blast.v2.csv”。从这个数据集,仅保留具有HLA-A/B/C等位基因的条目以及长度为8-15且无翻译后修饰的肽。这些样本的目标是定量(‘=’)或定性(‘<’/’>’),如在MHCflurry-1.2[3]中所建议的。MHCflurry精选数据集中的定性条目代表阳性高(<100nm)、阳性中等(<1000nm)、阳性低(<5000nm)或阴性(>5000nm)定性值。此外,还利用了细胞表面肽呈递数据集,所述细胞表面肽呈递数据集由如经由肽洗脱实验和质谱测定法从以下几个来源确定的细胞表面呈递肽组成:Sarkizova等人[4]数据集,其使用质谱法分析从95个HLA-A、HLA-B、HLA-C和HLA-G单等位基因细胞系中洗脱的>185,000个肽。质谱命中如在使用质谱法的MHCflurry精选数据集中所标识,其中相关样本是通过‘measurement_source’列中的‘质谱’值标识的。这包含保藏在IEDB[1]或System MHC Atlas[5]中或由Abelin等人[6]公布的226,684个经MS鉴定的配体。

[0174] 另外,使用了在哈钦森癌症研究中心(Fred Hutchinson Cancer Center)通过稳定转染HEK293细胞以表达与 β -2-微球蛋白共价连接的分泌的人类HLA分子而获得的细胞表面呈递数据。然后,制备细胞上清液以捕获分泌的MHC-肽复合物并通过质谱法进行分析。报道了来自纯化复合物的肽。数据文件可在压缩的“RolandPeptidePresentation Data.zip”文件中获得。这些数据源均由被发现在与特定HLA等位基因缔合的MHC I类分子上呈递的肽序列组成。对于我们的最终呈递数据集(“mass_spec_data.multiple_context.blast.allele_supertypes.v3.csv”),将来自所有提到的包含HLA-A/B/C等位基因和肽源蛋白的来源的样本合并,以便提取侧翼区并用于负采样方法。接近重复的“延伸”样本也被滤除,因为仅在任一边缘处具有单个附加氨基酸的具有相同等位基因和相同肽的样本被认为是重复的,这是由质谱测量不准确性引起的。将免疫原性评估数据中出现的样本(肽-MHC对)也从亲和力数据集和呈递数据集中滤除,以验证评估集在训练阶段中完全“保留”并且根本不被模型看到。

[0175] 在训练新模型之前,将这些数据集(亲和力和呈递)随机分成训练分割和验证分割。对于每个等位基因,我们从具有第*i*个等位基因的结合亲和力样本中随机采样 N_{bai} ,并从

具有第 i 个等位基因的呈递样本中随机采样 N_{pi} 肽。其中 $N_{bai} = \min(0.25 * |\text{独特亲和肽}|_i, 100)$ 且 $N_{pi} = \min(0.25 * |\text{独特呈递肽}|_i, 100)$ 。

[0176] 对于每个等位基因,这些采样肽(来自两个数据集的采样肽的并集)被认为是保留的验证集,并且包含这些肽的所有样本均被从训练集去除并且仅用于验证。如果验证损失在 N 个连续的时期(epoch)中没有改善,则使用验证集来确定模型训练的提前停止。所有实验中均设置 $N=20$ 。当训练多个相似模型时,出于集成/模型选择的目的,我们使用不同的训练验证分割。

[0177] 人鼻病毒(HRV)数据被包括在我们的数据中,作为来自应用于由HRV-1A、HRV-B、HRV-C嵌合体构建的1600个HRV 15聚体的ICS的阴性数据(如Fischer等人,(2007),Nature Medicine,13,100-106中所述)。具体地,使用从ICS结果中随机采样的阴性来进行免疫原性评估,并作为免疫原性评估数据的一部分进行测试。其余HRV样本被视为阴性呈递样本,并且它们是在训练期间随机采样的。

[0178] 实施例1.2:免疫原性评估数据

[0179] 为了验证对于肽-MHC结合亲和力和细胞表面肽呈递,如何良好地将来自BigMHC 1.0预测转化为预测T细胞免疫原性,创建T细胞免疫原性数据集,测试并验证BigMHC机器学习模型。使用来自HIV分子免疫学数据库中的CTL/CD8+表位总结表的报告的肽-MHC对和HCV免疫学数据库中的报告的CTL表位总结表。这些表提供了经实验验证的HIV/HCV CTL/CD8+表位。

[0180] 所有这些样本都属于阳性免疫原性类别。为了获得还具有阴性免疫原性的肽-MHC样本,我们手动审查了阳性对起源于的实验。已发现,虽然没有报告阴性发现,但它们中的一些可以从阳性结果重建。具体地,许多实验测试了给定的肽集合和给定的HLA等位基因集合的所有可能的肽-MHC对组合,这被称为矩阵方法。根据阳性发现,我们能够提取报告的肽和等位基因,并得出结论,至少(存在我们可能会错过的附加阴性,因为如果肽/等位基因在所有可能的配对中都是阴性的,则将不会报告所述肽/等位基因)在实验中测试了它们中所有可能的肽-MHC配对组合。鉴于此测试肽-MHC对的列表,可以得出的结论是这个列表中未报告为阳性的任何对实际上都是阴性的。审查了31个最大实验(具有最大的测试样本量,|独特的等位基因| $\times$ |独特的肽|),以验证是否使用了矩阵方法,其中18个最大实验使用了这种方法。从使用其他方法的实验,我们无法推断出阴性样本。

[0181] 对于剩余的较小实验,假设使用矩阵方法,并且提取所有阴性值,除非样本在不同的实验中被报告为阳性,在这种情况下我们假设所述样本是阳性的。仅保留HLA-A/B/C等位基因和长度为8-15的肽。

[0182] 除了上述数据之外,我们还添加了IEDB中报告的阳性免疫原性样本(Vita等人,(2019),Nucleic acids research,47,D339-D343.)和附加的随机采样的HRV阴性样本。对HRV阴性值进行采样

[0183] 对于具有相对低的阳性:性样本比率的等位基因。我们认为理想的比率为1:100,并且在可能的情况下(我们没有所有等位基因的HRV样本)对HRV阴性值进行采样,直到达到大约这个比率。按照这种平衡规程,我们过滤掉了比率小于1:5的所有等位基因样本。

[0184] 按照这个规程,获得了2,985个阳性样本对和68,469个阴性样本对,覆盖110个HLA等位基因和1,416个独特的肽。

[0185] 这个免疫原性数据集被分成两个集合—用于调谐我们模型的超参数并做出附加的配置选择的验证集6;以及测试集7,我们对所述测试集执行最终基准测试。

[0186] 实施例1.3:数据分析

[0187] 为了更好地可视化和理解数据的分布,绘制了下图。

[0188] 肽长度:图2B示出了跨所有三个数据集的肽长度分布。

[0189] 肽多样性:两个肽之间的相似性被称为具有最佳可能比对的重叠氨基酸的数量。在图2A中,对于每个相似性阈值,计算没有相似性大于给定阈值的配偶体肽的肽的比率。对于这种分析,仅考虑独特的肽,忽略每个数据集中重复的肽。具体地,亲和力数据集由总共158,001个样本中的35,467个独特肽(22.45%)组成,呈递数据集由总共384,812个样本中的265,236个独特肽(68.93%)组成,并且免疫原性数据集由总共71,474个样本中的1416个独特肽(1.98%)组成。

[0190] 目标分布:目标的分布以及它们之间的不平衡在不同的数据集之间是不同的:结合亲和力肽-MHC结合亲和力数据集由定量和定性目标的混合物组成。图3示出了结合亲和力靶标的高水平、定性分布。呈递—细胞表面肽呈递数据集是二进制的,并且我们的数据集仅由阳性样本组成。在训练器件,在每个时期开始时,我们应用阴性样本挖掘来针对每个阳性样本生成对应的阴性样本。免疫原性—T细胞免疫原性数据集是二进制的,但其高度不平衡,并且由2,985个阳性肽-MHC对和68,469个阴性肽-MHC对组成。

[0191] 每个超类型的样本分布—为了更好地可视化和理解数据集中的底层等位基因分布(这很难在每个等位基因的基础上可视化),图4示出了我们的数据集样本中HLA等位基因超类型的分布。应用由Sidney等人[7]确定的HLA超类型分类。

[0192] 每个HLA等位基因的样本分布—除了HLA等位基因超类型分布之外,图4还示出了数据集的肽-MHC样本中每个肽-MHC样本中HLA等位基因的分布。

[0193] 实施例1.4:训练期间阴性挖掘数据

[0194] 细胞表面肽呈递数据仅由“阳性”样本组成,但它们不提供训练二进制呈递分类器所需的阴性样本(其不能在细胞表面上呈递)。因此,为了训练此类分类器,采用以下用于训练期间概率阴性挖掘的策略:

[0195] HLA等位基因改组.给定由肽和对应的HLA等位基因组成的阳性样本的情况下,通过随机采样不属于阳性等位基因超类型的不同等位基因来替换给定等位基因。应用由Sidney等人[7]确定的HLA超类型分类,所述HLA超类型分类将每个HLA等位基因指派给一个或两个HLA超类型。这种分类留下了很少的未分类的HLA等位基因。将未分类的等位基因根据对应的HLA-A/B/C组映射到以下3个附加超类型类别中:“未分类的-A”、“未分类的-B”和“未分类的-C”,并以与其他超类型类别类似的方式处理这些组。

[0196] 肽改组.在给定由肽和对应的HLA等位基因组成的阳性样本的情况下,用来自肽源蛋白的具有相同长度的随机采样的氨基酸子序列替换给定的肽。此外,按照MHCflurry-1.6[8]中的方法,亲和力训练数据集也被扩充为包括从氨基酸数据分布中采样的具有定性弱亲和力目标(>20,000nM)的随机肽。这些随机肽的长度是以这样的方式确定的:对于每种等位基因,强制每个肽长度具有相同数量的非结合数据点。

[0197] HRV负采样.我们从阴性HRV数据中随机采样(排除用于免疫原性评估的样本)。这里的基本假设是,在大多数情况下,阴性免疫原性将由阴性细胞表面呈递引起,因此这可以

在训练期间假设(尽管存在具有阳性呈递和阴性免疫原性的情况)。

[0198] 实施例1.5:跨任务目标推断

[0199] 为了联合、多任务地进行训练,使用来自结合亲和力数据集和呈递数据集两者的所有训练样本。然而,对于每个样本,只有来自其起源于的对应数据集的单个已知目标(而不是两个任务)是已知的。为了缓解这个问题(并能够更好地利用多任务训练),通过假设在细胞表面上呈递的样本(阳性呈递)也将具有高结合亲和力值并且具有不良结合亲和力值的样本将不被呈递(阴性呈递),来推断从每个任务到另一个任务的目标。具体地,对于每个阳性呈递样本,我们推断出定性的高亲和力目标(<500nM),而对于具有不良的结合亲和力测量值的样本(>5000nM),推断出阴性呈递目标。其余的“缺失目标”(我们无法推断的目标)在训练期间通过指派零样本权重(仅用于缺失目标的任务)而被简单地忽略(掩蔽掉)。

[0200] 实施例1.6:自蒸馏

[0201] 使用BigMHC预测因子,我们提取了各种样本的结合亲和力和呈递估计,并将这些样本及其对应的“弱”标签添加到训练数据集中。我们在以下两种情况下执行了这种自蒸馏过程:

[0202] 1.多等位基因质谱数据.我们使用了如在MHCflurry-2.0中所述的MULTI-ALLELIC OLD数据集,其含有>200K的阳性质谱命中。我们利用BigMHC-1.3.1预测来确定等位基因中的哪个等位基因是导致命中的等位基因。首先,我们从我们的呈递训练数据中过滤掉任何具有已知阳性呈递者的多等位基因命中。其次,我们选择具有最高呈递概率的等位基因,并且仅当呈递概率高于某个阈值(0.5)并且结合亲和力低于某个阈值(5000nM)时才保留该样本。

[0203] 2.阳性呈递者.对于我们的训练数据中的每个具有未知结合亲和力的阳性呈递者(在执行上述步骤后),我们基于BigMHC-1.3.1预测来估计结合亲和力。我们将预测结合亲和力低于5000nM的所有样本添加到结合亲和力训练数据中。

[0204] 实施例2.1:序列表示

[0205] 每个HLA等位基因由49个氨基酸的伪序列表示,如在MHCflurry-1.4中所用。这种伪序列编码使用如通过跨物种的大量MHC I类等位基因的多重序列比对确定的49个选定位置处的氨基酸。所述HLA等位基因伪序列表示可在“allele_sequences.csv”文件中得到。此外,利用来自O'Donnell等人[3]的肽填充和编码方法,以使用被设计为保留与MHC进行最重要的稳定化接触的残基的定位的固定长度编码来表示氨基酸长度为8-15的肽。这些“锚定位置”出现在大多数等位基因的肽的开头或末端。肽表示为长度为15的序列,其中缺失的残基用“X”字符填充,实际上是第21个氨基酸。肽中的第一个和最后四个残基映射到所述表示中的第一个和最后四个位置。中间的七个残基根据需要填充:8聚体将所有中间位置保留为X,而15聚体则填充所有位置。以这种方式,最有可能包含锚定残基的位置被一致地映射到所述表示中的相同位置。肽的侧翼区也被编码,考虑成每侧5个氨基酸,所述氨基酸分别拼接在所编码的肽的边缘处。

[0206] 与使用基于BLOSUM62替换矩阵的固定氨基酸嵌入的O'Donnell等人[3]相反,使用可训练的嵌入层,所述可训练的嵌入层是与我们的神经网络的其余部分以端到端的方式联合训练的。这个嵌入层将编码肽或等位基因伪序列中的每个氨基酸编码到16维向量中。

[0207] 实施例2.2:侧翼区

[0208] 对于每个肽序列,鉴定了这个序列是UniProt数据集[9]中存在的较长蛋白质序列的子集的所有实例。搜索了以下3个UniProt文件:(1)UniProt人类蛋白质组数据集,“UP000005640_9606.fasta”; (2)完整的UniProtKB/Swiss-Prot数据集,“uniprot_sprot.fasta”; (3)UniProtKB/Swiss-Prot数据集的表示所有带注释的剪接变体的附加序列,“uniprot_sprot_varsplic.fasta”;和(4)来自netMHCpan-4.0免疫原性数据集的附加序列,CD8(“CD8_epitopes_netMHCpan.fas”)和CD4(“CD4_epitopes_netMHCIIpan.fsa”)均被下载。

[0209] 较长序列中的每个较长序列被称为“亲本序列”。每个肽与一个或多个长度为10的“侧翼区”缔合,所述“侧翼区”是每个所述肽的亲本序列中紧邻所述肽序列之前的5个氨基酸和紧邻所述肽序列之后的5个氨基酸。将侧翼区的所有独特组合保存到文件中,并定义与每种肽的独特序列数成反比的每种独特组合的权重。这些权重在训练期间用作样本权重,以便让网络从所有可能的变型中学习,但不会更加重视具有大量变化的肽。对于没有精确匹配的肽,我们利用BLAST10找到最接近匹配的肽,并使用所述肽的对应“亲本序列”来提取相关的侧翼区。

[0210] 实施例2.3:自监督预训练

[0211] 我们利用大型蛋白质数据库来预训练我们的模型并学习良好的初始序列表示。受BERT预训练启发,使用来自Uniparc数据库的25M蛋白质的子集,我们针对以下两个任务训练肽转换器模型:

[0212] 1.掩码语言建模.我们随机选择0.15个被视为“被掩蔽”的令牌并训练令牌分类头,所述令牌分类头尝试使用交叉熵损失来预测原始令牌(基于所有其他令牌)。对于模型输入,“被掩蔽的”令牌有时被随机替换(10%),有时保持不变(10%),有时被掩蔽令牌替换(80%)。

[0213] 2.下一个肽预测.在预训练阶段中,我们的输入序列是两个肽序列的拼接(与主训练阶段中肽和等位基因序列的拼接相反)。所述序列经由特殊的分离令牌(<SEP>)进行分离并具有不同的分段索引和嵌入(所述分段序列是网络的附加输入,简单地指示每个令牌是属于第一序列、第二序列还是特殊令牌。然后将所述分段索引的嵌入添加到令牌中并定位嵌入)。我们在<CLS>令牌的输出上训练分类器,以预测第2个肽是否是蛋白质中下一个出现的肽(在第1个肽之后)。这些肽来自人类蛋白质的两个连续的相同长度的肽,或者是从不同蛋白质随机采样的。

[0214] 实施例2.3训练目标和多任务损失

[0215] 我们的神经网络经联合训练以预测肽-MHC结合亲和力和细胞表面肽呈递。正如O'Donnell等人[3]所做的那样,我们利用由LBA-MSE表示的均方误差(MSE)损失函数的变体,由此与不等式(>)或(<)相关联的测量仅在不等式被违反时才有助于所述损失,以用于处理数据集中的定量和定性肽-MHC结合亲和力测量两者。确切的表达式如方案1和方案2中所概述。

[0216] 对于细胞表面肽呈递的二进制分类任务,我们利用由LP-FL表示的焦点损失[10],所述焦点损失是更加强调分类不佳的样本的标准二进制交叉熵损失的加权扩展。确切的表达式在方案3中概述(下面再次再现)。

$$[0217] \quad L_{P-FL} = -\frac{1}{N} \sum_{i=1}^N (1 - p_{ti})^\gamma \log p_{ti},$$

[0218] 其中正确 P_{ti} 的预测概率为:

$$[0219] \quad p_{ti} = \begin{cases} \hat{Y}_i & , Y_i = 1 \\ 1 - \hat{Y}_i & , Y_i = 0 \end{cases}$$

[0220] 方案3

[0221] 在此, γ 是实参数, 我们将其设置为1。在二进制情况下, $Y_i \in \{0, 1\}$ 是地面实况标签, 并且 $\hat{Y}_i \in [0, 1]$ 是第 i 个样本的预测呈递概率。Lin 等人 [10] 表明焦点损失对于处理数据不平衡是有效的; 并且 Mukhoti 等人 [11] 的最近工作还表明, 与标准交叉熵损失相比, 它可以产生更好的经校准网络。总体目标函数是肽-MHC 结合亲和力的 MSE 变体和细胞表面肽呈递的二进制焦点损失的线性组合, $L = \alpha L_{BA-MSE} + (1 - \alpha) L_{P-FL}$ 。

[0222] 在训练期间, 我们还应用了 in 实施例 1.4 中提到的负采样策略。它们在每个时期开始时应用于呈递和结合亲和力任务两者。验证集的阴性样本在第一个时期开始时生成一次, 并在整个训练过程中固定。在可能的情况下, 还利用了从每项任务到另一项任务的推断目标。对于每个损失项, 还应用了样本权重机制, 以基于侧翼区变化的量为样本赋予不同的权重, 并且还掩蔽掉丢失目标的样本 (我们仅从目标未知的特定任务中掩蔽掉它们)。如实施例 1.5 中所解释, 具有推断目标的样本对于这两项任务具有相同的样本权重。

[0223] 对于细胞表面肽呈递预测, 我们用二进制交叉熵损失训练二进制分类器: 由 L_{P-BCE} 表示的呈递目标是:

$$[0224] \quad L_{P-BCE} = -\sum_i (Y_i \log \hat{Y}_i + (1 - Y_i) \log(1 - \hat{Y}_i)).$$

[0225] 方案9

[0226] 总体目标函数是上述训练目标的线性组合, 具有预定义的重量损失系数:

$$[0227] \quad L = \alpha L_{P-BCE} + \beta L_{BA-MSE} + \gamma L_{A-MLM}.$$

[0228] 方案10

[0229] 在训练期间, 我们还对我们使用所述四种方法随机采样的每个阴性样本应用了负采样策略。我们还定义了阴性: 阳性比率的超参数 N_{ratio} , 我们用它来确定每个阳性呈递样本要采样多少阴性样本。对于每个采样的阴性样本, 我们指派了 $1/N_{ratio}$ 的样本权重, 由此有效地确保 L_{P-BCE} 目标是在平衡数据上训练的 (阳性训练样本和阴性训练样本的总体样本权重相等)。此外, 在主要训练阶段期间, 我们利用掩蔽语言建模辅助目标, 其中仅从肽的侧翼区随机选择掩蔽令牌。这个目标仅应用于天然肽序列, 并且因此对于某些类型的采样阴性样本被忽略, 在所述采样阴性样本中肽 (例如随机采样的氨基酸序列) 或其背景 (例如 HRV 阴性样本) 是“合成的”。我们用 L_{A-MLM} 表示这种辅助损失。

[0230] 实施例 2.4 校准

[0231] Guo 等人 [12] 的最近工作表明现代神经网络校准不佳。校准性质对我们来说非常重要, 因为在我们的排名逻辑管线中, 我们使用概率计算来基于预测的呈递概率做出决策。具体地, 在疫苗设计管线中, 载给定受试者的 6 个 HLA 等位基因的情况下, 我们应用以下概率计算以基于模型对每个单一 HLA 等位基因的预测来估计受试者等位基因的总呈递概率:

$$[0232] \quad P_{\text{呈递}}(\text{肽} | \text{等位基因}_1, \dots, \text{等位基因}_6) = 1 - \prod_{i=1}^6 (1 - P_{\text{呈递}}(\text{肽} | \text{等位基因}_i))$$

[0233] 方案4

[0234] 如果我们的预测因子被正确校准,则这些类型的计算将会得到增强。因此,我们通过将低次多项式与校准曲线拟合来进一步在验证集上校准网络的呈递预测。所有多项式系数都被限制为正以获得单调递增函数。对于这个任务,使用了Lasso线性回归。为了获得经良好校准的呈递概率,这个步骤至关重要,所述经良好校准的呈递概率稍后将用于我们的推理管线中,然而,由于校准步骤是单调的,因此它不会影响单个等位基因的肽的排名。在我们的疫苗设计管线中,我们使用所述经校准的呈递预测作为免疫原性概率估计的最佳指标。

[0235] 实施例2.5模型架构

[0236] A. 模型架构1

[0237] 这个模型的架构由3个主要组成部分组成:序列处理、肽-MHC结合亲和力预测,之后是细胞表面肽呈递预测。我们模型的输入是肽一级序列,包括如上所述的侧翼区,以及长度为49个氨基酸的HLA等位基因伪序列。首先将肽序列编码到固定长度的载体中,然后使用共享的d维氨基酸嵌入层对所有氨基酸序列进行编码。然后将嵌入序列展平,以产生每个肽、等位基因和侧翼区的向量表示。对于肽-MHC结合亲和力预测组成部分,我们拼接肽和HLA等位基因表示,并应用2个大小为512和256的致密层,每个致密层之后是指数线性单元(ELU)激活和暂退概率为 $p=0.5$ 的暂退层。这个组成部分的输出被称为“亲和力表示”。然后使用附加的致密线性层来输出预测的结合亲和力logit。最后,对于细胞表面肽呈递预测组成部分,首先将肽、侧翼区和HLA等位基因表示拼接到单个向量中。然后利用一系列2个完全连接层,其配置与肽-MHC结合亲和力预测组成部分中类似,但是上面的“亲和力表示”的输出也被拼接。这个组成部分的输出被称为“呈递表示”。然后,添加附加线性致密层来预测细胞表面肽呈递概率logit。

[0238] B. 模型架构1

[0239] 所述模型由三个组成部分构成:1.序列嵌入;2.自注意力转换器层;和3.预测头。

[0240] 1.序列嵌入—模型接收以下2个序列作为输入:1)氨基酸序列(等位基因伪序列和肽与侧翼区的拼接)。2)分段标识符序列,所述分段标识符序列“通知”模型每个氨基酸属于哪个分段(等位基因、肽、背景和特殊令牌)。图12展示了令牌和分段输入序列的组成。所述序列的每个令牌由aa嵌入、所学习的位置嵌入和分段嵌入相加表示。

$$[0241] \quad x = \text{aa_embed}(x) + \text{pos_embed}(x) + \text{seg_embed}(\text{seg})$$

[0242] 方案11

[0243] 2.自注意力转换器层。嵌入序列是使用12个连续的转换器层处理的,每个转换器层包含多头自注意力模块之后是前馈模块(由2个线性层以及其间的GELU激活构成),如图13所示。在每个模块的开始处应用层归一化,并且在残差连接之前在每个组成部分的结束处应用速率 $p=0.1$ 的残差暂退。

[0244] 3.预测头-所述模型包含以下3个预测头:

[0245] 掩蔽语言建模:将序列的每个位置处的最终表示馈送到LM预测头,所述LM预测头

由线性层(具有gelu激活之后是层归一化)和(至令牌词汇量大小的)附加线性投影构成,所述附加线性投影具有与令牌嵌入矩阵相关联的权重和附加学习偏差。

[0246] 结合亲和力:<CLS>令牌位置处的最终表示被馈送到由以下构成的预测头中:线性+GELU+层归一化+暂退+线性。

[0247] 呈递:将在<CLS>令牌位置处与单个结合亲和力logit拼接的最终表示馈送到由以下组成的预测头中:线性+GELU+层归一化+暂退+线性

[0248] 实施例2.6模型集成

[0249] 使用上述神经网络架构来训练以下2种类型的模型:(1)泛等位基因,其中包含所有训练数据;和(2)等位基因特异性,其中训练数据按HLA等位基因划分。使用足够的训练数据为每个HLA等位基因训练单独的HLA等位基因特异性模型(利用数据集中至少有1K结合亲和力和1K呈递样本的标准)。

[0250] 两种类型具有完全相同的架构并且经过类似的训练。它们的区别仅在于它们的训练数据:推理时支持的等位基因的列表,及它们的权重初始化。虽然泛模型是随机初始化的,但是等位基因特异性模型是根据经最佳训练的泛模型微调的。在推理时,利用经训练的模型的集成,所述集成对每个样本在支持所述样本的等位基因的所有模型上的预测进行平均。

[0251] 实施例2.7模型选择

[0252] 早期实验表明,模型的子集,或者甚至单个模型,往往比许多模型的集成表现得更好。因此,我们开发了以下模型选择规程:1.对于每个给定的模型配置(具有足够训练数据的针对每个等位基因的泛模型和等位基因特异性模型),我们仅使用不同的折叠(训练-验证分割)来训练10个具有精确训练设置的模型。在训练完成后,对于每个模型配置,我们基于我们的评估方案,使用验证免疫原性数据分割,从所述10个经训练的模型中选择单个表现最佳的模型。我们应用每等位基因分层模型选择,其中对于每个等位基因(具有用于性能验证的免疫原性数据),我们从以下选项中选择最佳可能的配置:a)仅使用等位基因特异性模型;b)仅使用泛模型;c)使用泛模型+等位基因特异性模型的集成并对它们的预测进行平均。

[0253] 实施例2.8评估

[0254] 为了评估模型性能,利用T细胞免疫原性数据来检查给定的经训练模型如何良好地管理以将阳性免疫原性肽-HLA等位基因对排名成高于阴性非免疫原性肽-HLA等位基因对。具体地,大约排名前20的肽是令人感兴趣的,因为这大致是制造给定疫苗所需的肽的量。在给定免疫原性验证/测试集的情况喜爱,可以在所有样本上提取模型的预测。考虑到这一点,利用聚焦于排名前K的项目的3个常见排名度量:Precision@K、nDCGK和倒数排名。还利用在先前的工作(诸如O'Donnell等人[13])中使用的阳性预测值度量。对于每个等位基因,在免疫原性验证/测试集中,基于预测的细胞表面肽呈递概率或预测的结合亲和力得分对所有对应的肽单独进行排名,并计算每个等位基因的Precision@K、nDCGK、倒数排名和阳性预测值度量。

[0255] 在单独计算每个等位基因的这些度量之后,我们通过应用加权平均将它们跨所有HLA等位基因进行汇总,其中每个等位基因按照其在美国群体中的频率来加权。我们使用了美国中的4个最大族群的等位基因频率[11]。将通过这些文件得出的频率进一步按照美国

群体中每个种族的频率进行加权[12]。具体地,对于欧洲白种人使用0.54,对于西班牙裔使用0.22,对于非洲裔美国人使用0.17,并且对于亚洲人使用0.07。按群体中等位基因频率进行加权的动机是为了使所述度量捕获受试者的比率(或至少更好地与受试者的比率相关),这种方法可理论上能够提供帮助。免疫原性评估集中HLA等位基因的基于群体的频率权重绘制在图7中。最终度量,加权的Precision@K (WP@K)、加权的nDCG_K (WnDCG_K)、加权的倒数排名 (WRR) 和加权的阳性预测值 (WPPV) 由下式给出:

$$[0256] \quad WP@K = \sum_i \tilde{f}_i \frac{|\text{阳性样本} \cap \text{排名前K的样本}|}{K_i}$$

$$[0257] \quad WnDCG_K = \sum_i \tilde{f}_i \frac{DCG_K}{IDCG_K},$$

[0258] 其中

$$[0259] \quad DCG_K = \sum_{k=1}^K \frac{2^{rel_k} - 1}{\log(k+1)},$$

[0260] 方案5

[0261] rel_k 是第k个排名项的地面实况相关性,其在我们的情况下是二进制免疫原性标签。IDCG_K是参考项的基于地面实况的理想排名的DCG_K得分。

$$[0262] \quad WRR = \sum_i \frac{\tilde{f}_i}{\text{排名}_i}$$

[0263] 方案6

[0264] 其中,排名_i是第i个等位基因的经排名的肽中第一阳性样本的排名。

$$[0265] \quad WPPV = \sum_i \tilde{f}_i \frac{|\text{阳性样本} \cap \text{排名前} N_i \text{的样本}|}{N_i}$$

[0266] 方案7

[0267] 其中, N_i 是第i个等位基因的阳性样本的数量。在所有度量中,

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

[0268]

方案 8

[0269] 是基于美国人口统计的HLA等位基因频率权重,并且它们的值在[0,1]范围内(越高越好)。在实践中, $K=20$ 用于 W_nDCG_K 度量,并且对于WP@K,使用 $K_i = \min(20, |\text{阳性样本}|_i)$,因为我们的评估数据中一些等位基因具有少于20个阳性样本。

[0270] PPV度量虽然信息丰富,但并不聚焦于排名前几的肽,这是我们最感兴趣的区域。倒数排名度量使用经排名的肽中第一阳性项目的排名,排名越高,则给出的得分越高。然而,这个度量忽略了排名靠前的项目中是否存在附加阳性肽。由于我们希望确保我们具有尽可能多的阳性结果(因为并非所有呈递的肽都是免疫原性的并产生所需的免疫应答),所以在我们设计的疫苗中,它并不理想。Precision@K度量仅捕获排名前K的项目中的阳性率。然而,它没有考虑它们在前K中的实际排名(意指一个排名第一的阳性项目对比一个排

名第K的阳性项目将获得完全相同的得分,这并不理想)。nDCGK度量考虑了前K内的阳性率及其对应的排名。如果阳性数量低于20,则按照理想排名的DCG的归一化因子也会减少限制K的需要,并且还会产生在[0,1]范围内的值。然而,与precision@K度量相比,这些得分仍然不太可解释/直观。

[0271] 实施例3.1结果

[0272] 除非另有明确说明,否则我们用ADAM优化器,使用为0.001的学习率和为256的批量大小来训练所有模型。我们设置 $a=0.5$ 作为损失项系数,赋予两个损失项相等的权重。在20个时期后应用提前停止,但在保留的验证集上评估的我们的验证损失中没有观察到改善。应用为 $1e-9$ 的小L1正则化因子和具有0.5暂退率的暂退两者。对于所有模型,都应用负采样方法和目标推断。对于等位基因特异性模型,在我们不想用来自不同超类型的等位基因替换来自阳性样本的等位基因(因为我们只对特定等位基因的样本感兴趣)的情况下,我们反转实现方式并保留(在过滤来自其他等位基因的所有数据之前)具有来自其他超类型(除了我们正在训练的特定等位基因的超类型之外)的等位基因的阳性样本存储库,以及来自这个集合的随机采样的阳性样本,与此同时用我们正训练的当前等位基因替换“外部”等位基因。对于最终预测因子,训练以下模型类型的集合:1.泛模型;和2.等位基因特异性模型。此类模型仅针对具有足够训练数据的等位基因子集进行训练。具体地,训练具有至少1K结合亲和力和1K呈递训练样本的等位基因。

[0273] 对于每种类型,在不同的训练-验证分割上训练10个相似的模型,并选择免疫原性验证集上表现最佳的模型。应用分层模型选择来确定每个等位基因在推理期间使用哪个模型。

[0274] 为了验证多任务训练在我们的设置中是否有益,对损失权重系数 a 的效应进行了探索。执行几个泛模型3倍实验,其中除了这个超参数之外具有相同的设置,并报告了免疫原性测试分割的平均WP@K。绘制通过亲和力预测排名和通过呈递预测排名两者获取的结果以捕获完全效应。图7清楚地示出了 $a=0$ 和 $a=1$,它们分别对应于仅呈递训练和仅亲和力训练,从而导致与使用两个目标的加权组合进行联合训练相对应的中间值相比,导致较差的结果。

[0275] 为了验证我们的呈递概率是免疫原性估计的良好指标的假设,用每个箱的阳性免疫原性样本的比率绘制分箱的呈递预测的直方图。如图8A所示的正斜率证实了 $\frac{dP_{\text{免疫原性}}}{dP_{\text{呈递}}} > 0$,即,平均而言,增加肽的呈递概率会增加所述肽为免疫原性的机会。为了进行比较,我们还探索了关于结合亲和力预测的等效行为并在图8B中观察到了类似的图案。我们还比较了我们的模型与其他现有最先进技术预测因子之间的性能。

[0276] 具体地,我们将性能与来自O'Donnell等人[13]及其以前的版本的MHCflurry-2.0进行了比较。使用我们的评估方案和度量进行比较,所报告的数字是在免疫原性测试分割上计算的。从表1可以看出,我们的最佳预测因子显著优于MHCflurry-2.0预测因子和之前版本的所有变体。我们还可以看到,我们的单一泛模型优于等位基因特异性模型的集合(这是有道理的,因为它们不支持对所有等位基因的预测),但两者与分层模型选择的组合提供了附加的性能提升。

[0277] 表1. 评估产生了所提议的免疫原性基准。最佳预测因子与MHCflurry预测因子之间的性能比较。所报告的度量是:加权的Precision@K (WP@K)、加权的nDCGK (WnDCGK)、加权

的倒数排名 (WRR) 和加权的阳性预测值 (WPPV), 总而言之越高越好。

预测因子		$WP@K$	$WnDCG_{20}$	WRR	$WPPV$
[0278]	BigMHC-2.0 呈递预测因子	0.589	0.651	0.763	0.492
[0279]	亲和力预测因子	0.539	0.602	0.725	0.459
	BigMHC-1.3 呈递预测因子	0.535	0.591	0.744	0.418
	BigMHC-1.3 亲和力预测因子	0.466	0.526	0.703	0.396
	MHCflurry-2.0 呈递预测因子+侧翼	0.421	0.501	0.627	0.343
	MHCflurry-2.0 呈递预测因子(无侧翼)	0.424	0.5	0.659	0.348
	MHCflurry-2.0 亲和力预测因子	0.468	0.547	0.69	0.383

[0280] 我们另外比较了我们的最佳预测因子“泛+等位基因特异性-呈递预测因子”与 MHCflurry-2.0最佳预测因子“泛(+ms)-亲和力预测因子”之间的每等位基因性能(针对所有具有待评估的免疫原性数据的等位基因)。结果呈现于表2中。

[0281] 表2. 本文所述的最佳预测因子与MHCflurry-2.0基线(所建议的免疫原性基准)之间每等位基因的性能比较。所报告的度量是: Precision@K (P@K)、nDCG20、倒数排名 (RR) 和阳性预测值 (PPV), 均越高越好。

等位基因	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
[0282] 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

[0283]

等位基 因	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-	0.31	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
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

[0284]

等位基 因	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-	0.65	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	0.95 0	0.955	1.00 0	0.66 3	1.00 0	1.000	1.00 0	0.58 4	0.85 0	0.841	1.00 0	0.56 2
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
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

[0285]

等位基 因	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-	0.00	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

[0286] 实施例中引用的参考文献

[0287] 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 reseatch 2019,47,D339-D343.

[0288] 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.

[0289] 3.0'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.

[0290] 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.

[0291] 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 SystemMHC atlas project.Nucleic acids research 2018,46,D1237-D1247.

[0292] 6.Abelin,J.G.;Harjanto,D.;Malloy,M.;Suri,P.;Colson,T.;Goulding,S.P.;Creech,A.L.;Serrano,L.R.;Nasir,G.;Nasrullah,Y.;others.Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer epitope

prediction.Immunity 2019,51,766-779.

[0293] 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.

[0294] 8.O'Donnell,T.;Rubinsteyn,A.;Laserson,U.Improved predictive models for peptide presentation on MHC I.BioRxiv 2020.

[0295] 9.Consortium,U.UniProt:a worldwide hub of protein knowledge.Nucleic acids research 2019,47,D506-D515.

[0296] 10.Lin,T.Y.;Goyal,P.;Girshick,R.;He,K.;Dollár,P.Focal loss for dense object detection.Proceedings of the IEEE international conference on computer vision,2017,pp.2980-2988.

[0297] 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.

[0298] 12.Guo,C.;Pleiss,G.;Sun,Y.;Weinberger,K.Q.On calibration of modern neural networks.arXiv preprint arXiv:1706.04599 2017.

[0299] 13.O'Donnell,T.J.;Rubinsteyn,A.;Laserson,U.MHCflurry 2.0:Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing.Cell Systems 2020.

[0300] 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.

[0301] 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.

[0302] 16.Consortium,U.UniProt:a worldwide hub of protein knowledge.Nucleic acids research 2019,47,D506-D515.

[0303] 6.等同物

[0304] 对于本领域技术人员来说将显而易见的是,本文所述的本发明的方法的其他合适的修改和适应是显而易见的,并且可以在不脱离本公开或实施方案的范围的情况下使用合适的等同物进行。现在已经详细描述了某些组合物和方法,通过参考以下实施例将更清楚地理解相同的组合物和方法,引入这些实施例仅用于说明而非限制。

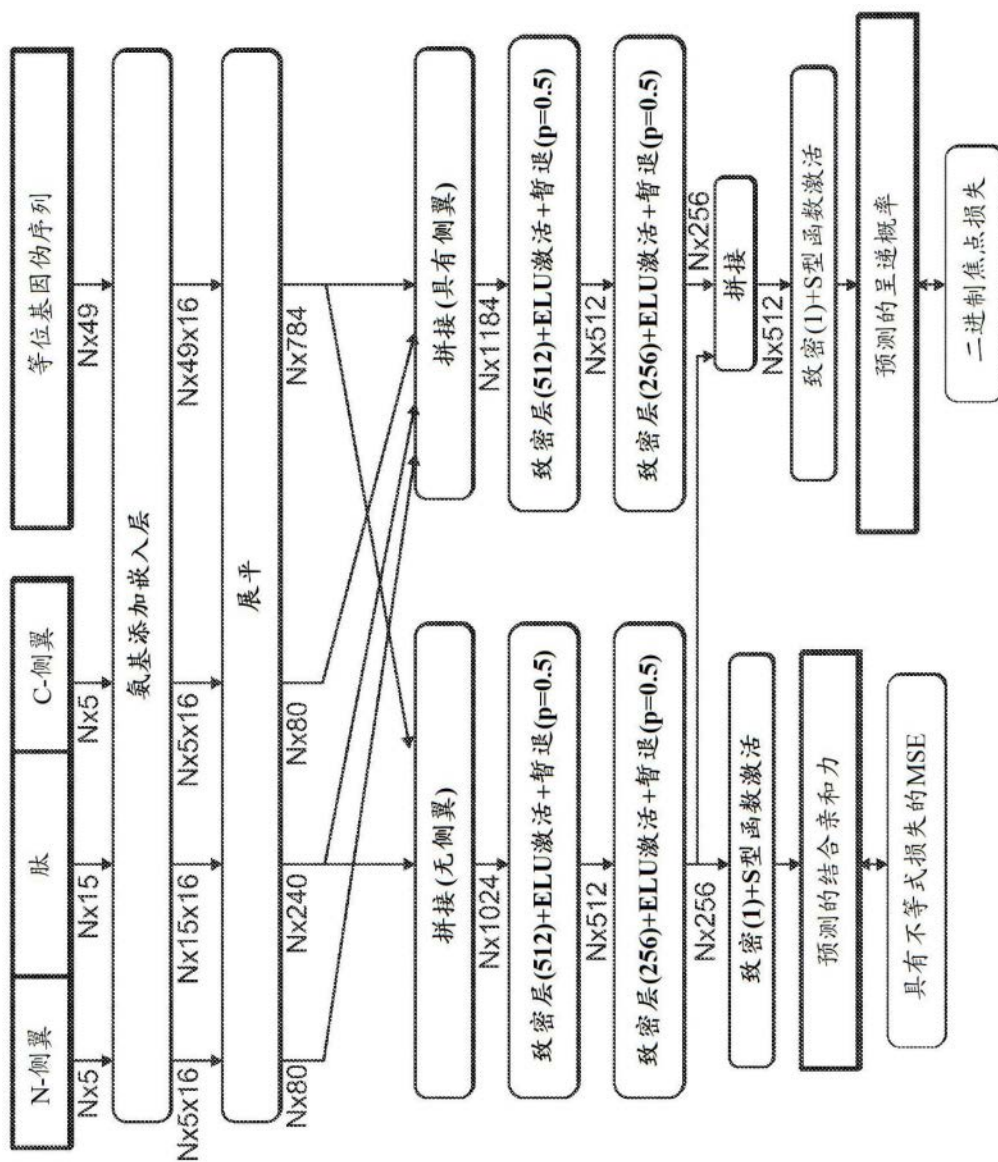


图1

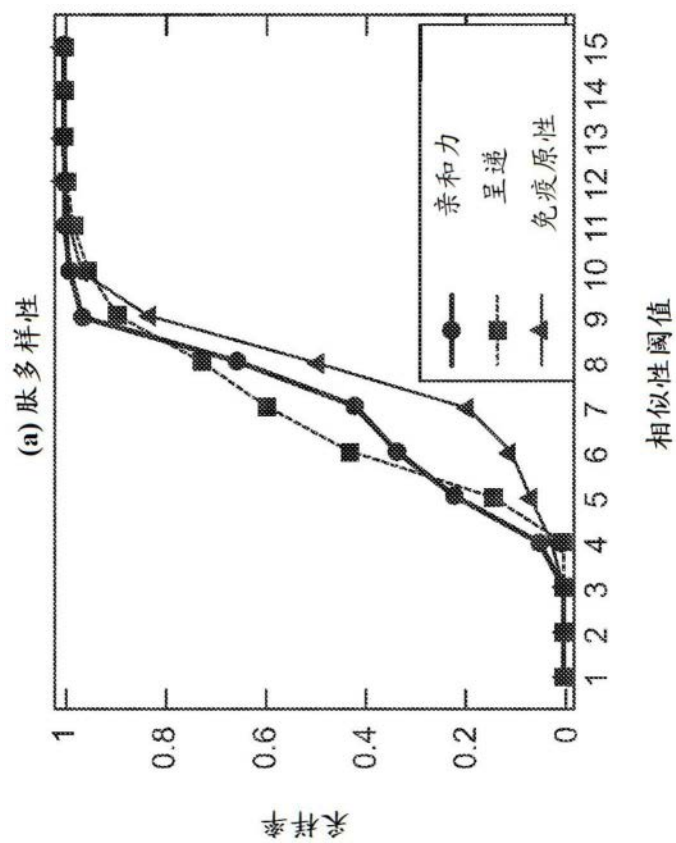


图2A

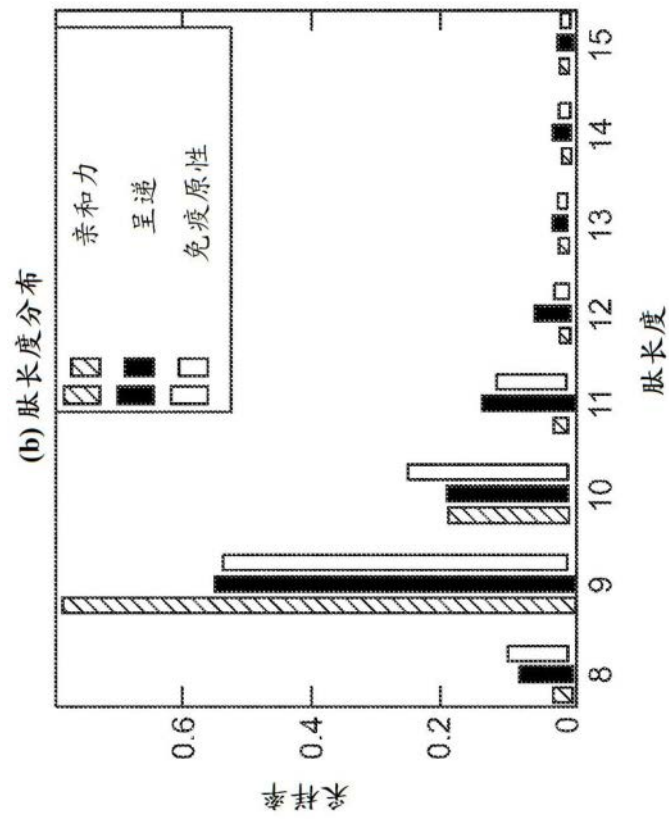


图2B

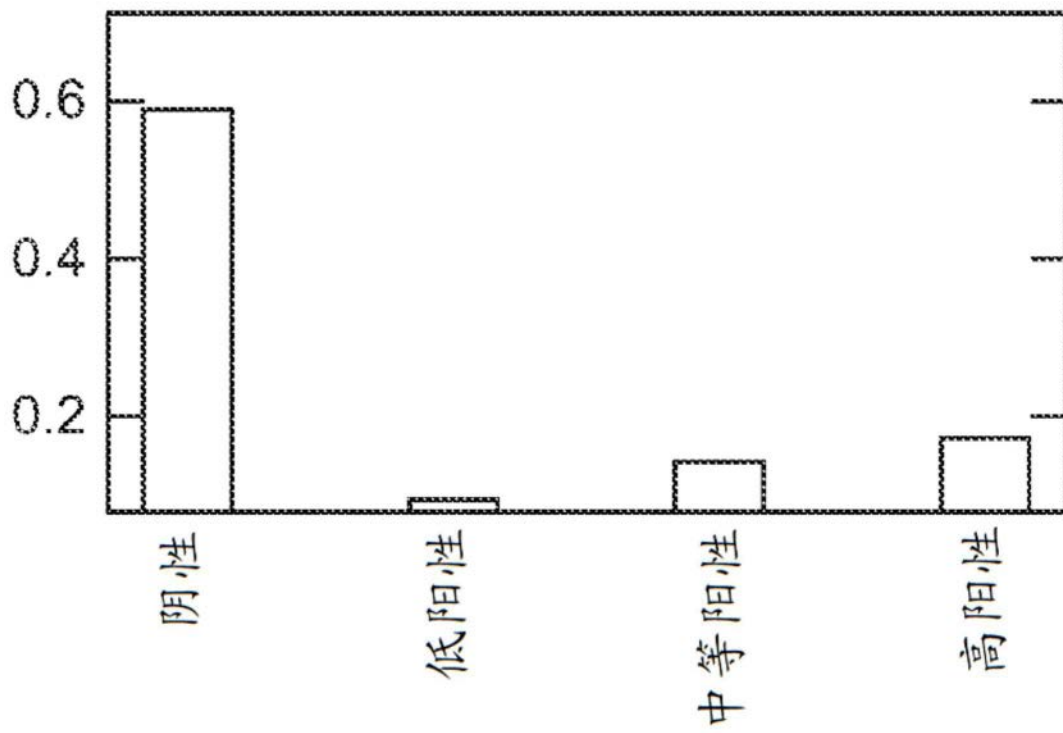


图3

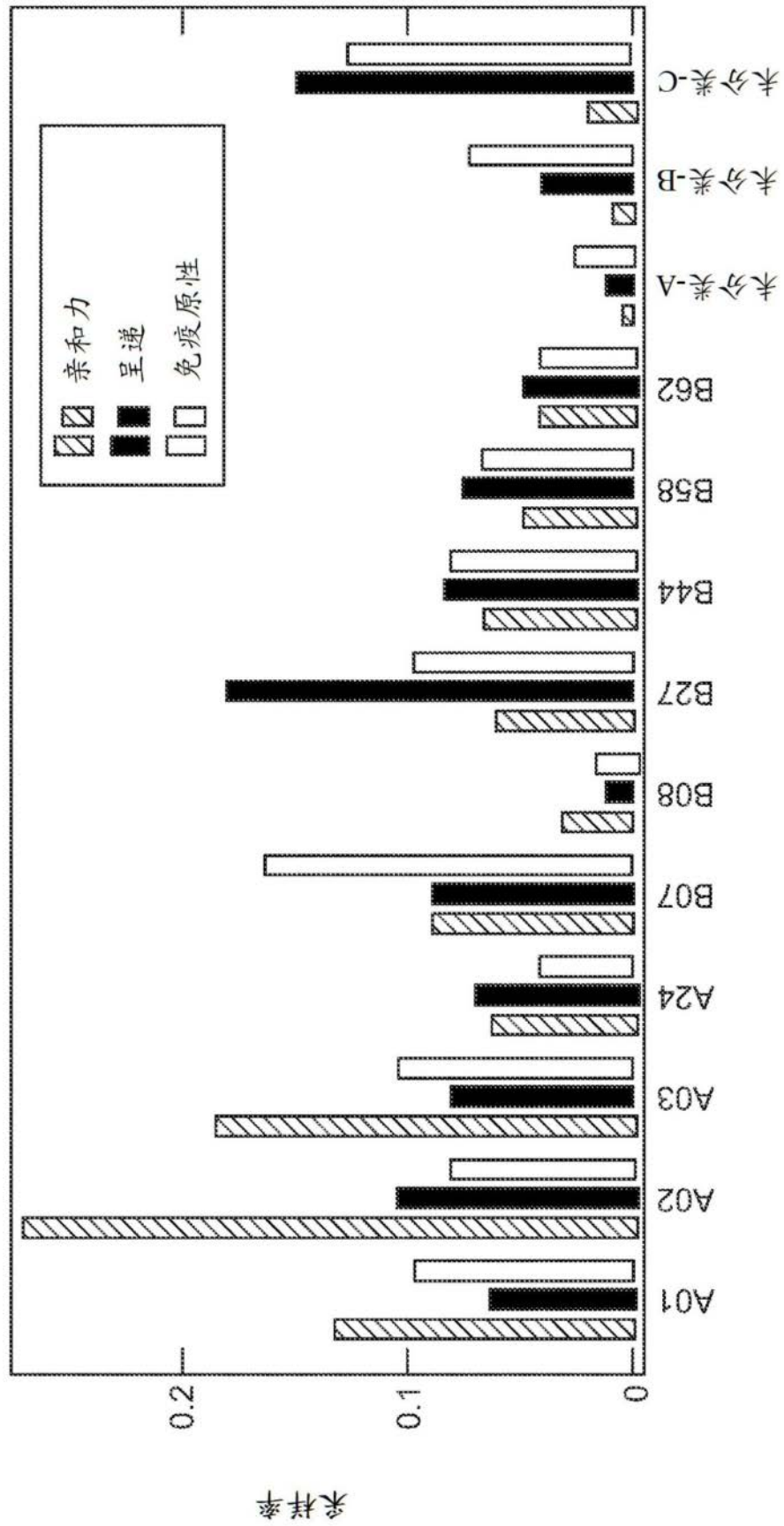


图4

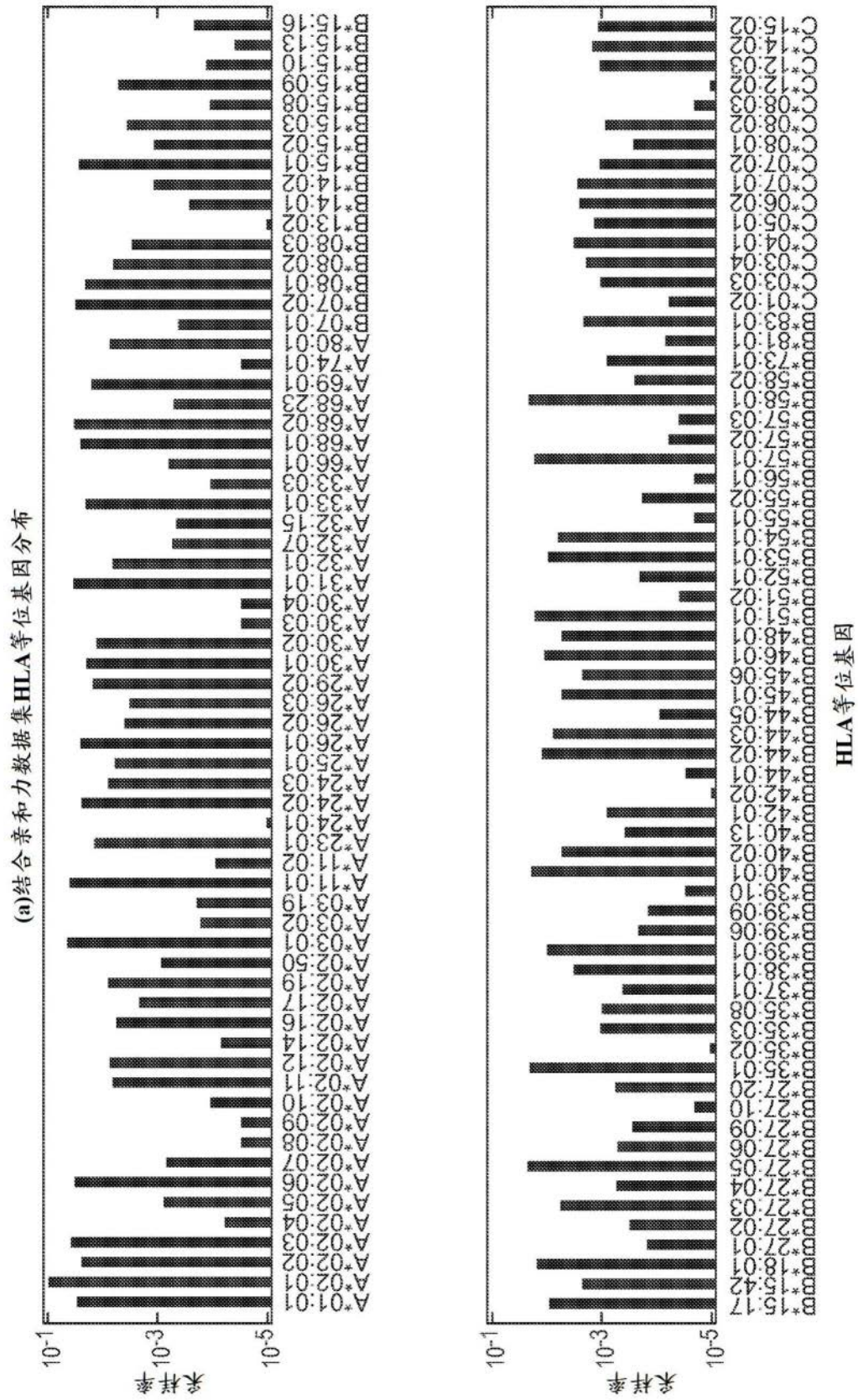


图5A

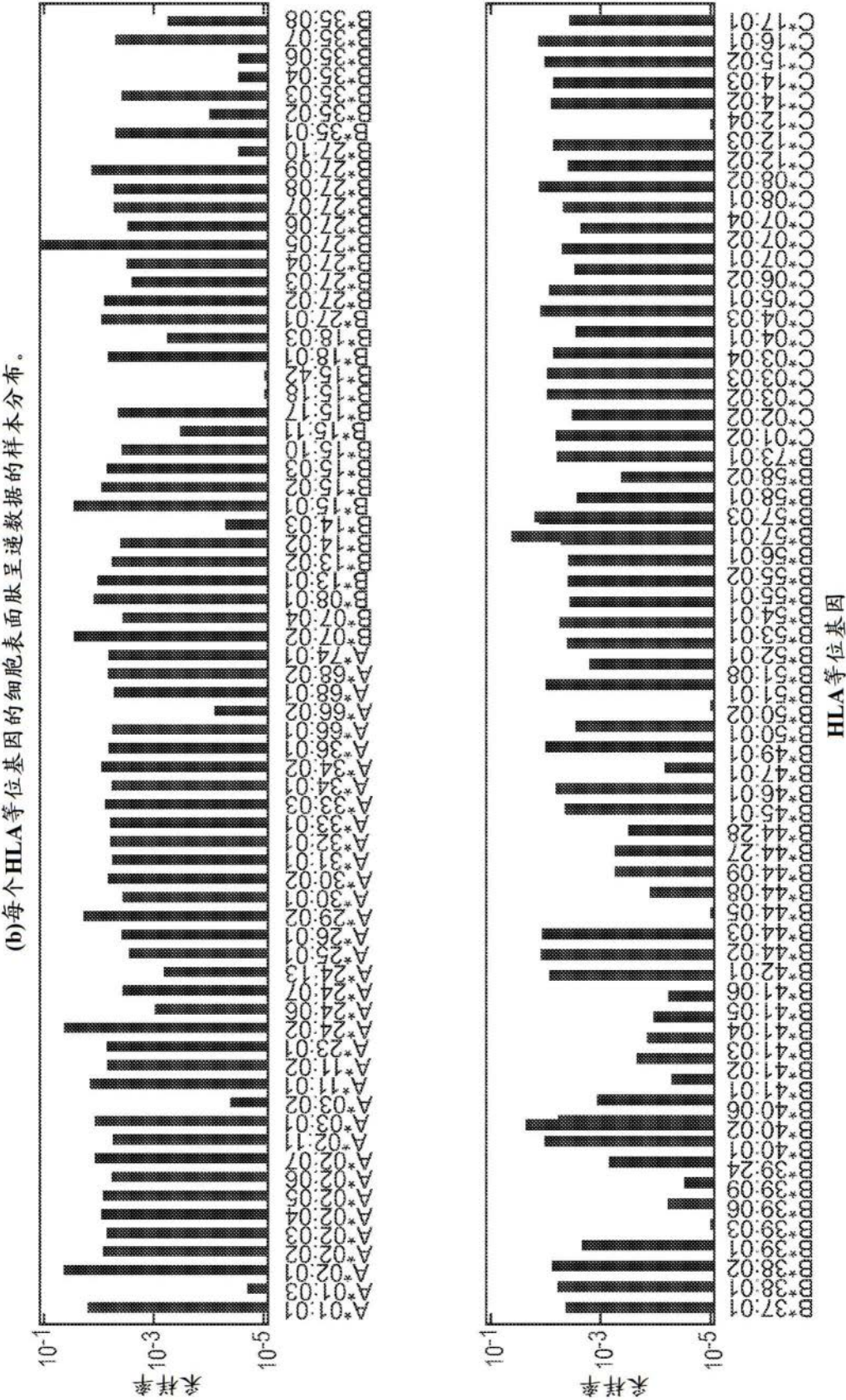


图5B

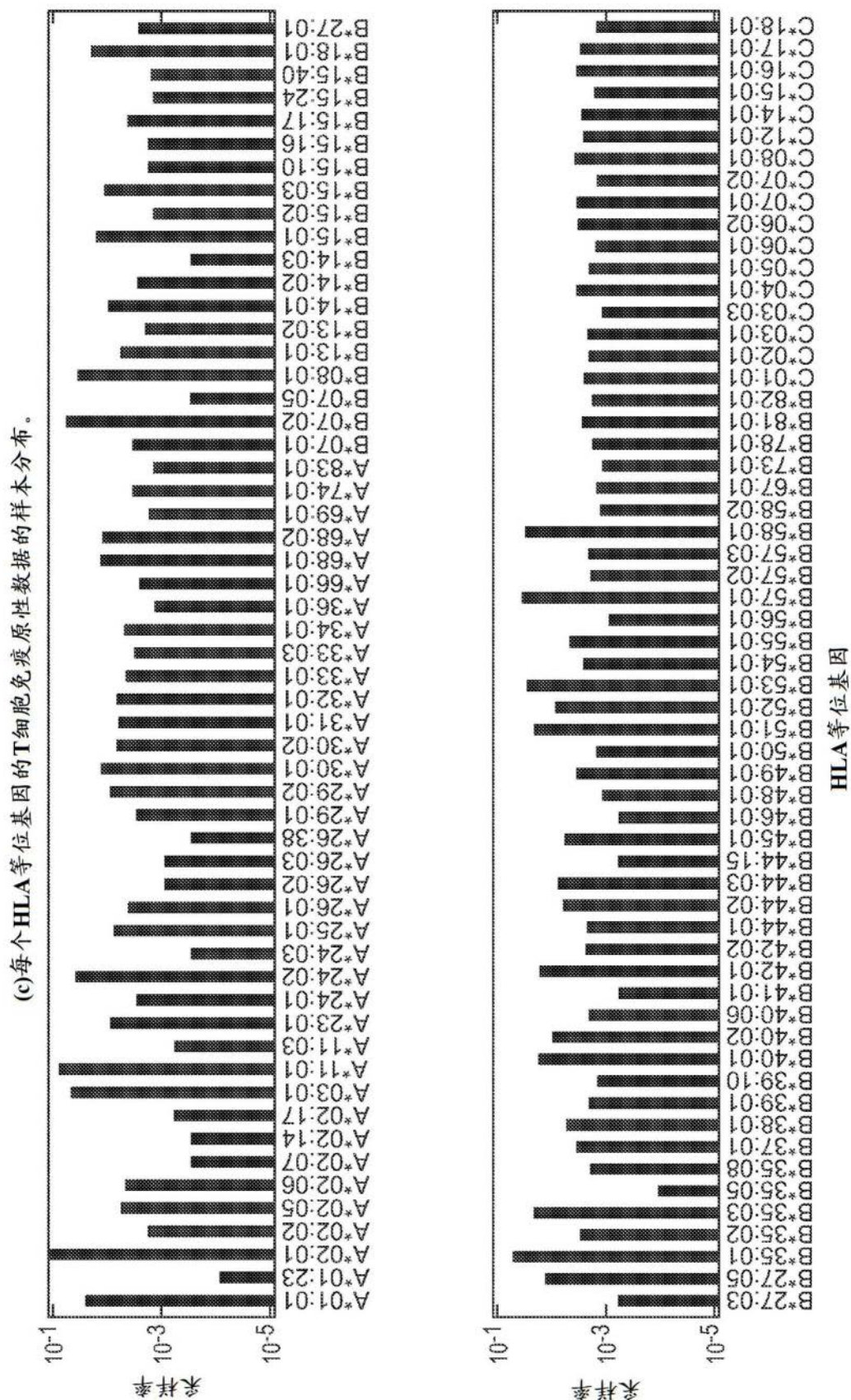


图5C

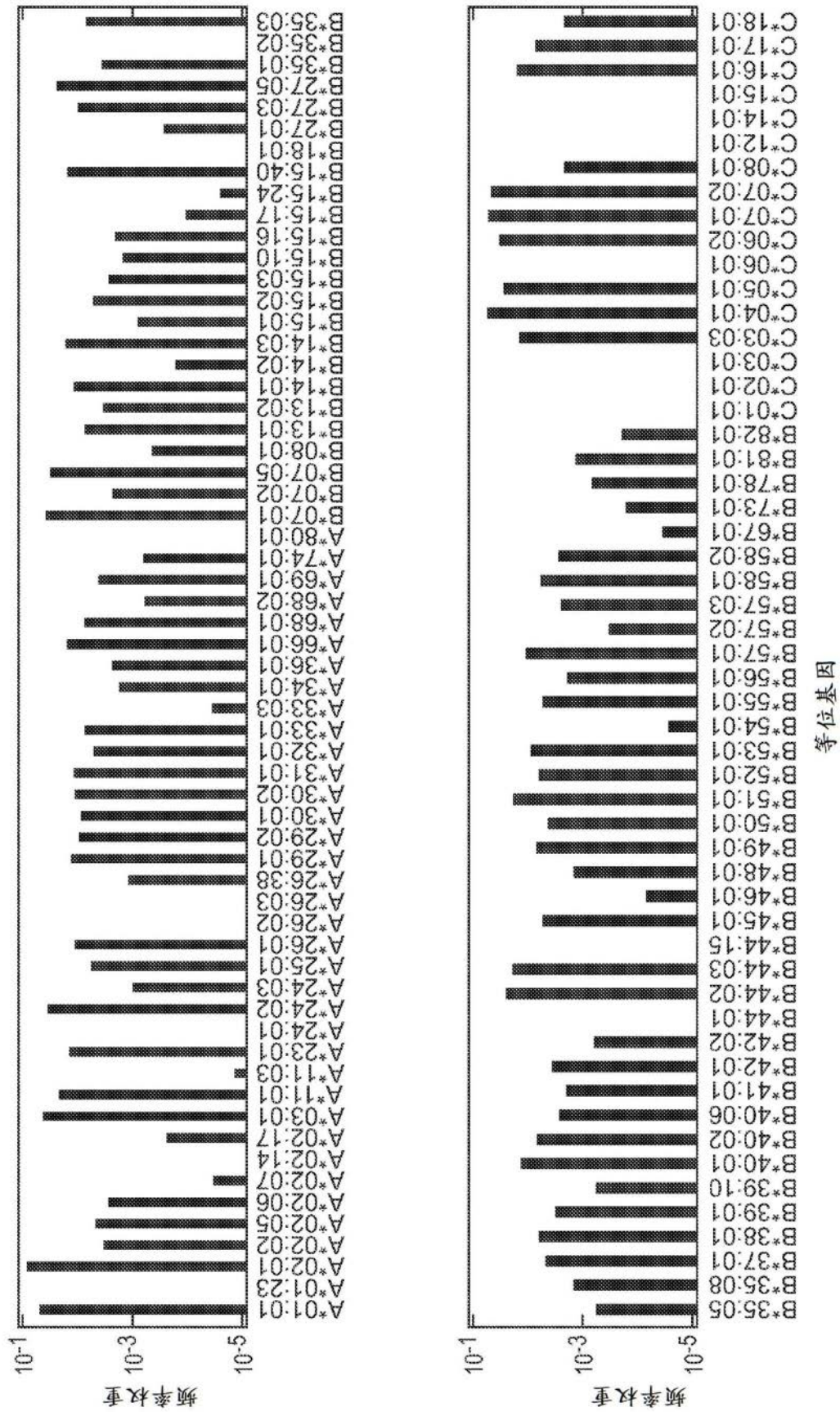


图6

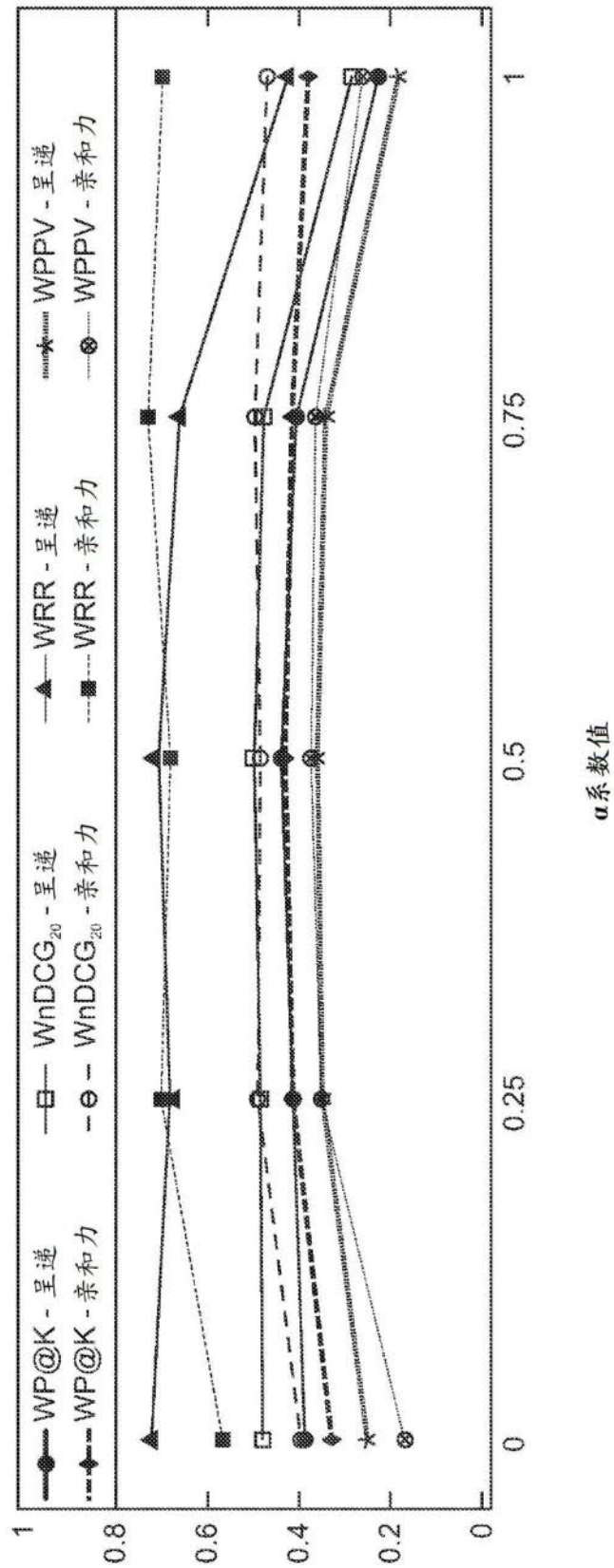


图7

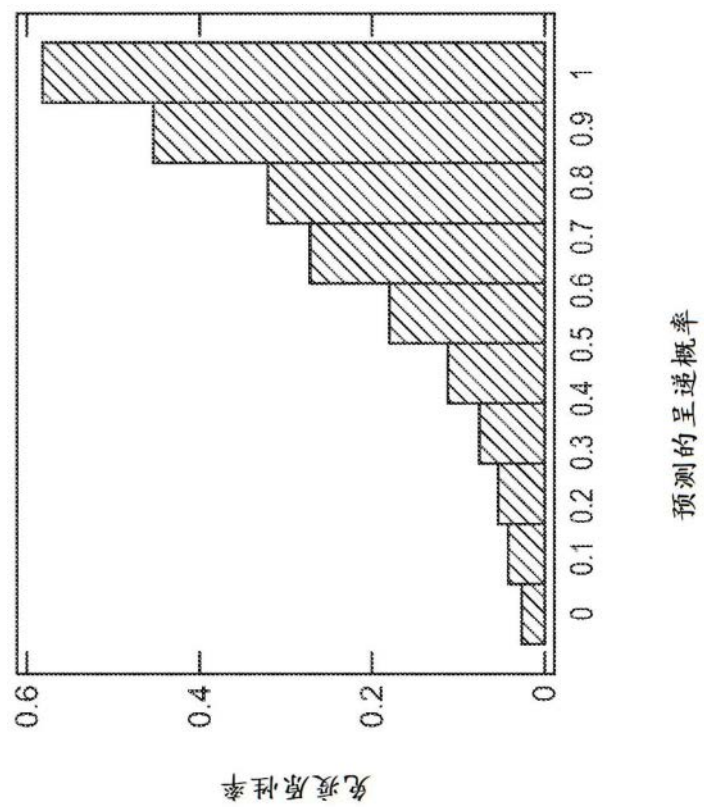


图8A

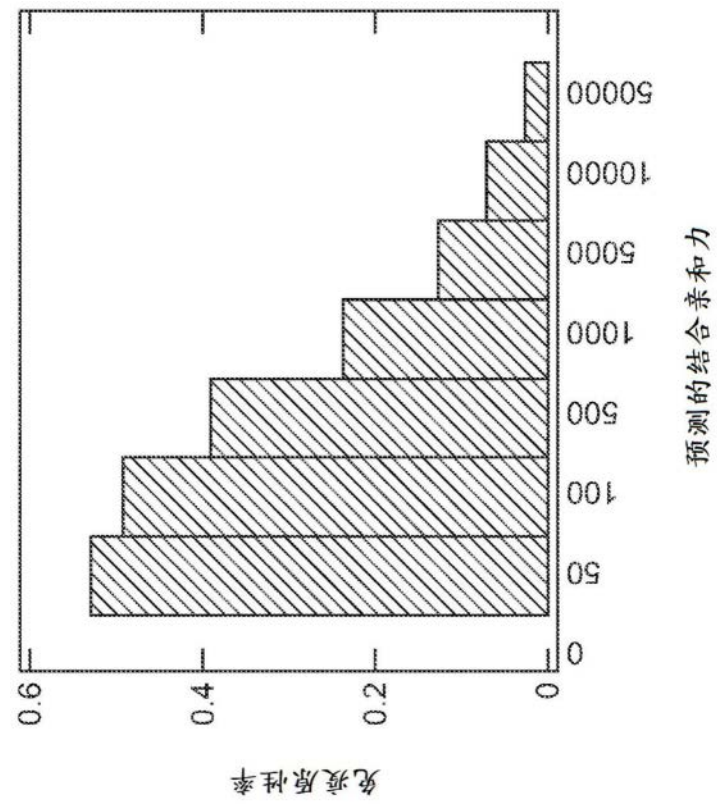


图8B

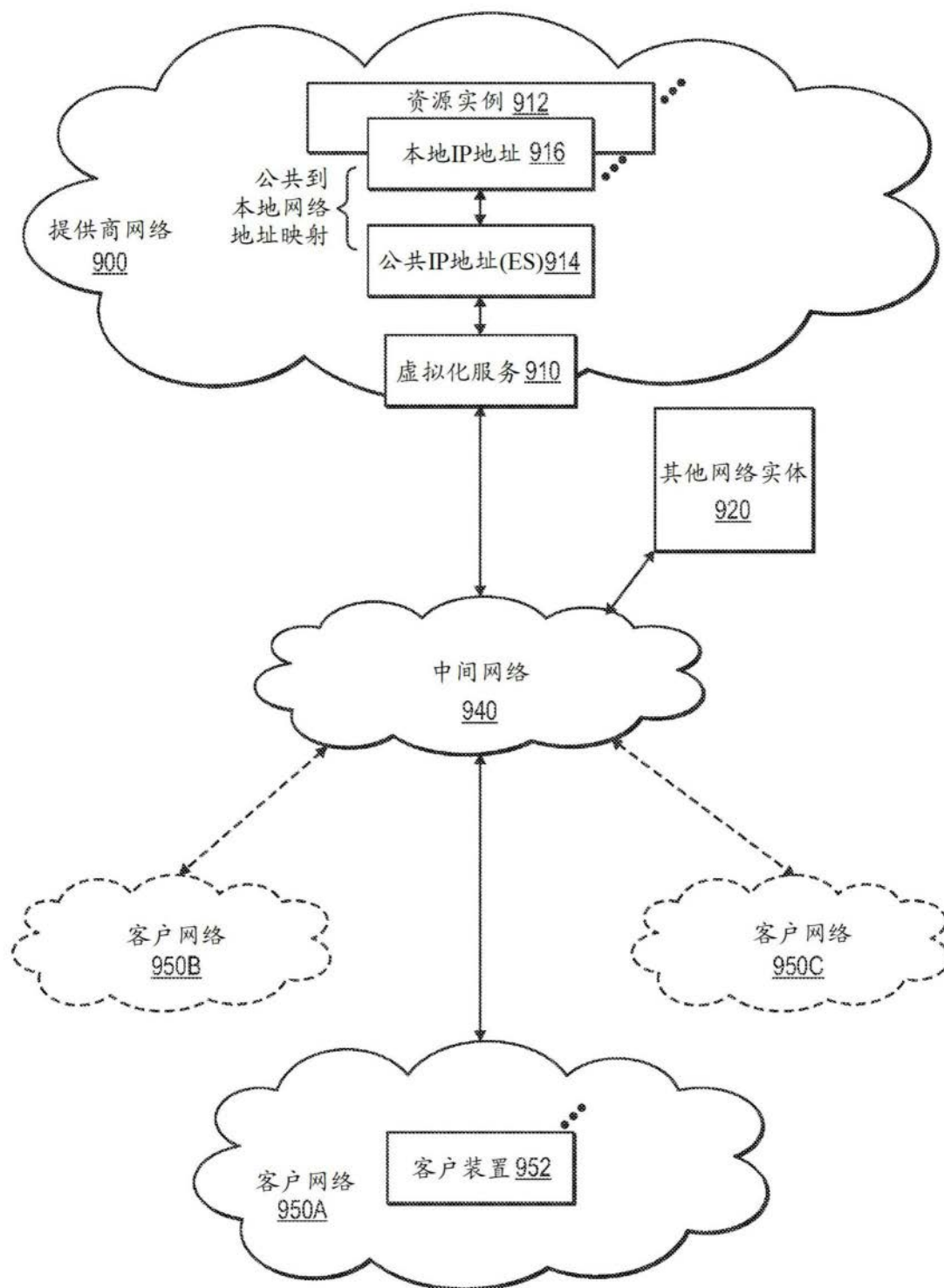


图9

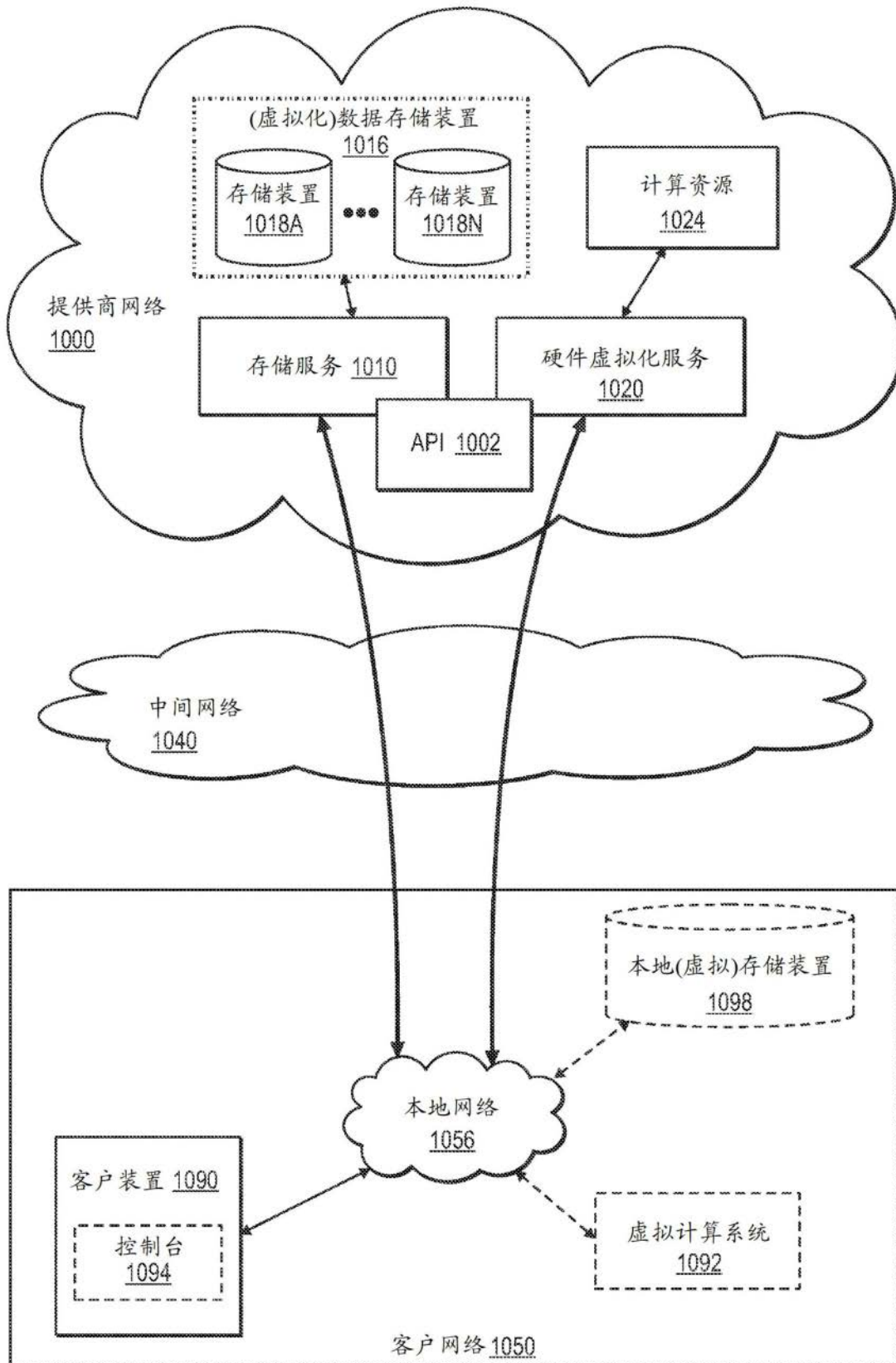


图10

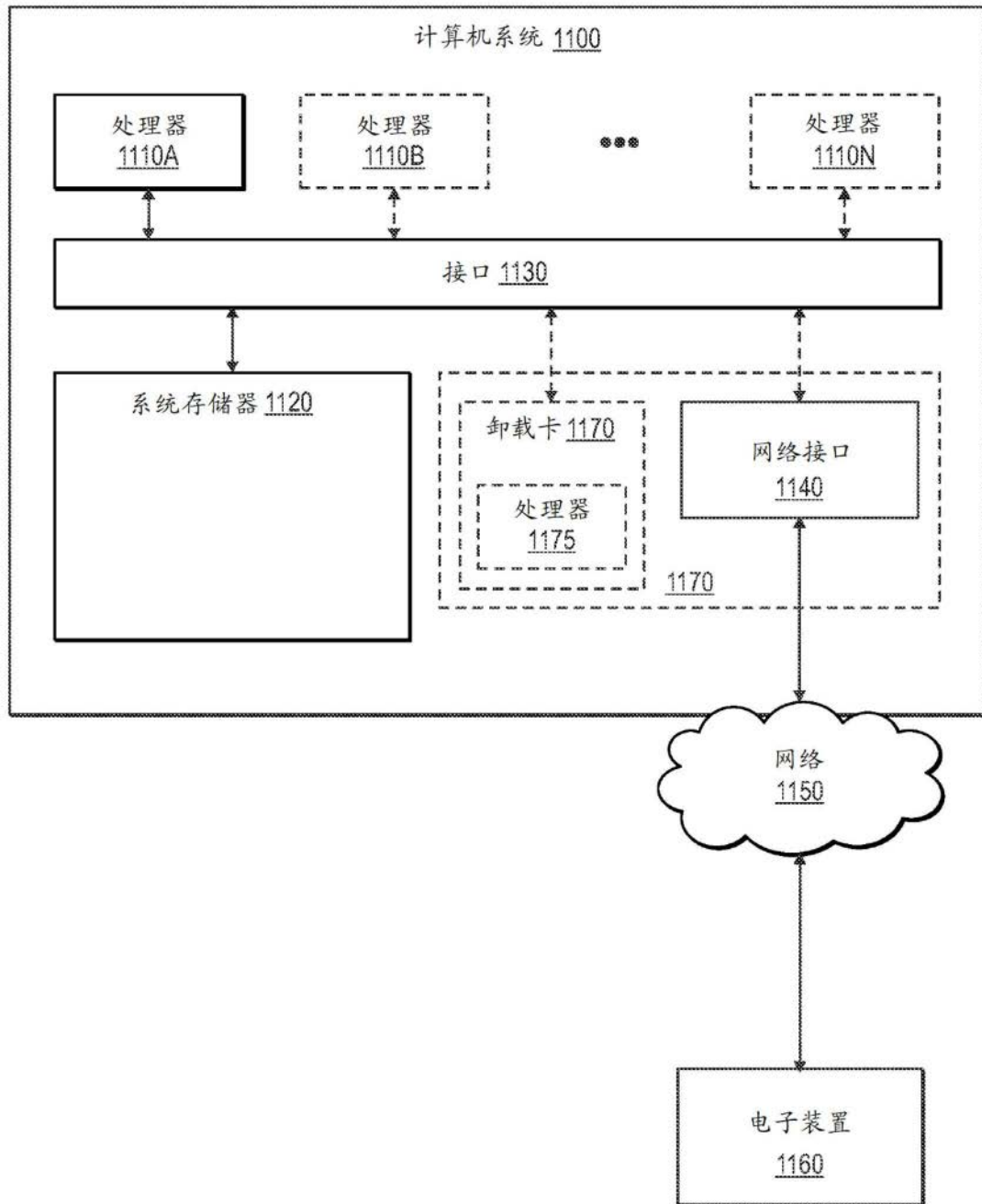


图11

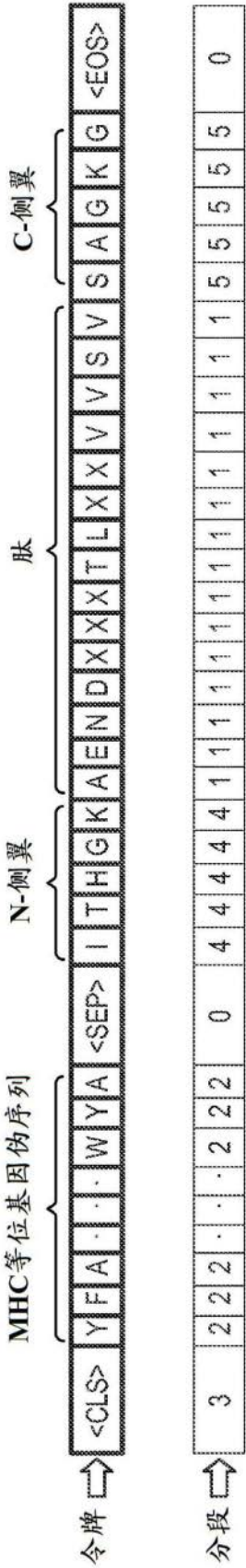


图12

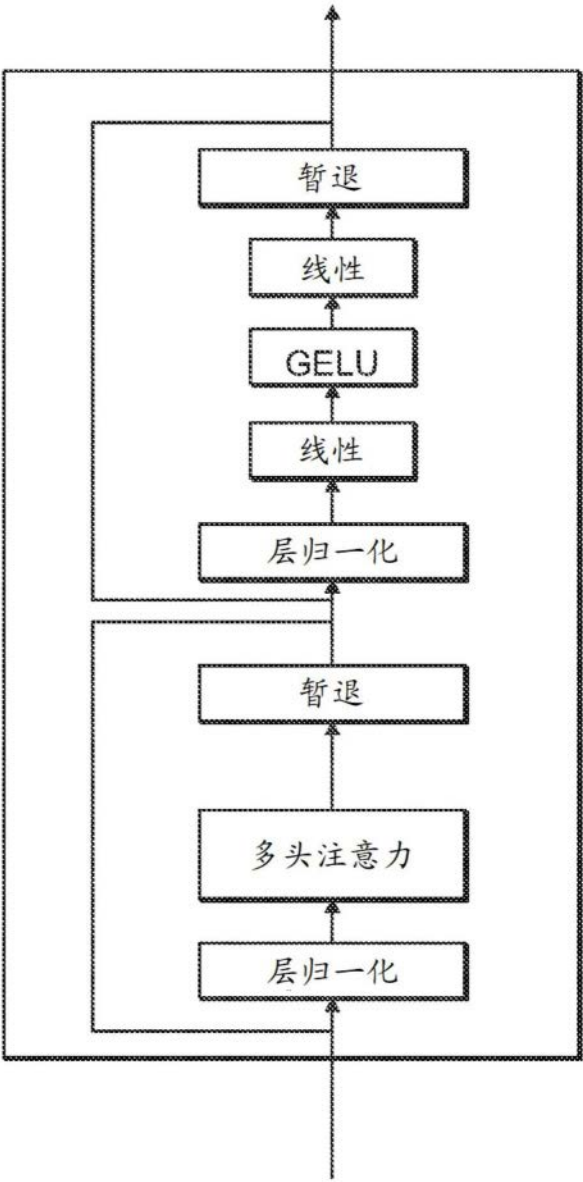


图13