(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111630602 A
(43)申请公布日 2020.09.04

(21)申请号 201880086681.4

(22)申请日 2018.11.21

(30)优先权数据
62/590,045 2017.11.22 US

(85)PCT国际申请进入国家阶段日
2020.07.15

(86)PCT国际申请的申请数据
PCT/US2018/062294 2018.11.21

(87)PCT国际申请的公布数据
WO2019/104203 EN 2019.05.31

(71)申请人 磨石肿瘤生物技术公司
地址 美国加利福尼亚州

(72)发明人 B·布里克-沙利文 T·F·鲍彻
R·耶冷斯凯 J·巴斯比

(74)专利代理机构 北京市金杜律师事务所 11256
代理人 陈文平 袁元

(51)Int.Cl.
G16B 20/50(2019.01)
G16B 40/00(2019.01)
G01N 33/68(2006.01)
G01N 33/574(2006.01)
C40B 30/04(2006.01)
C40B 40/10(2006.01)

权利要求书6页 说明书65页 附图23页

(54)发明名称

减少新抗原的接合表位呈递

(57)摘要

给定治疗性表位集合,设计盒序列来减小接合表位在患者中呈递的可能性。所述盒序列通过考虑跨越所述盒中一对治疗性表位之间的接合部的接合表位的呈递来设计。盒序列可基于各自与所述盒的接合部相关联的距离度量集合来设计。距离度量可指示将呈递跨越一对相邻表位之间的所述一个或多个接合表位的可能性。

CN 111630602 A

1.一种鉴别新抗原疫苗的盒序列的方法,其包括:

对于患者,从所述受试者的肿瘤细胞和正常细胞获得外显子组、转录组或全基因组肿瘤核苷酸测序数据中的至少一种,其中所述核苷酸测序数据被用于获得表示通过比较来自所述肿瘤细胞的核苷酸测序数据和来自所述正常细胞的核苷酸测序数据所鉴别的新抗原集合中每一者的肽序列的数据,其中每种新抗原的所述肽序列包含至少一个使其不同于从所述受试者的正常细胞鉴别的相应野生型亲本肽序列的变化,并且包含关于构成所述肽序列的多个氨基酸和所述肽序列中所述氨基酸的位置集合的信息;

使用计算机处理器将所述新抗原的所述肽序列输入机器学习的呈递模型中以生成所述新抗原集合的数字呈递可能性集合,所述集合中的每个呈递可能性表示相应新抗原由一个或多个MHC等位基因呈递在所述受试者的肿瘤细胞表面上的可能性,所述机器学习的呈递模型包括:

至少基于训练数据集鉴别的多个参数,所述训练数据集包含:

对于样品集合中的每个样品,标记,其通过测量结合至鉴别为存在于所述样品中的MHC等位基因集合中的至少一个MHC等位基因的肽的存在的质谱法获得;

对于每个所述样品,训练肽序列,其包含有关构成所述训练肽序列的多个氨基酸和所述训练肽序列中所述氨基酸的位置集合的信息;以及

一个函数,其表示作为输入接收的所述新抗原的所述肽序列与作为输出生成的所述呈递可能性之间的关系;

对于所述受试者,鉴别来自所述新抗原集合的新抗原治疗子集,所述新抗原治疗子集对应于具有高于预定阈值的呈递可能性的新抗原的预定数目;以及

对于所述受试者,鉴别包含串接的治疗性表位的序列的盒序列,所述治疗性表位各自包含所述新抗原治疗子集中相应新抗原的肽序列,其中所述盒序列基于跨越一对或多对相邻治疗性表位之间的相应接合部的一个或多个接合表位的呈递来鉴别。

2.如权利要求1所述的方法,其中所述一个或多个接合表位的所述呈递基于通过将所述一个或多个接合表位的序列输入所述机器学习的呈递模型中所生成的呈递可能性来确定。

3.如权利要求1所述的方法,其中所述一个或多个接合表位的所述呈递基于所述受试者的所述一个或多个接合表位与所述一个或多个MHC等位基因之间的结合亲和力预测来确定。

4.如权利要求1所述的方法,其中所述一个或多个接合表位的所述呈递基于所述一个或多个接合表位的结合稳定性预测来确定。

5.如权利要求1所述的方法,其中所述一个或多个接合表位包括与第一治疗性表位的序列和串接在所述第一治疗性表位之后的第二治疗性表位的序列重叠的接合表位。

6.如权利要求1所述的方法,其中连接子序列置于第一治疗性表位与串接在所述第一治疗性表位之后的第二治疗性表位之间,并且所述一个或多个接合表位包括与所述连接子序列重叠的接合表位。

7.如权利要求1所述的方法,其中鉴别所述盒序列包括:

对于每对排序的治疗性表位,确定跨越所述一对排序的治疗性表位之间的所述接合部的接合表位集合;以及

对于每对排序的治疗性表位，确定指示所述排序对的所述接合表位集合在所述受试者的所述一个或多个MHC等位基因上的呈递的距离度量。

8.如权利要求1所述的方法，其中鉴别所述盒序列包括：

生成对应于所述治疗性表位的不同序列的候选盒序列集合；

对于每个候选盒序列，基于所述候选盒序列中每对排序的治疗性表位的所述距离度量确定所述候选盒序列的呈递评分；以及

选择与具有低于预定阈值的呈递评分相关联的候选盒序列为所述新抗原疫苗的所述盒序列。

9.如权利要求8所述的方法，其中随机生成所述候选盒序列集合。

10.如权利要求7所述的方法，其中鉴别所述盒序列还包括：

解出以下优化问题中x_{km}的值：

$$\min_{x} \sum_{k=1}^{v+1} \sum_{k\neq m, m=1}^{v+1} P_{km} \cdot x_{km}$$

$$\sum_{k=1}^{v+1} x_{km} = 1, \quad m = 1, 2, \ldots, v+1$$

$$\sum_{m=1}^{v+1} x_{km} = 1, \quad k = 1, 2, \ldots, v+1$$

$x_{kk}=0, k=1,2,\ldots,v+1$

$$\text{out}(S) \geq 1, \quad S \subset E, 2 \leq |S| \leq |V|/2$$

其中v对应于新抗原的预定数目，k对应于治疗性表位并且m对应于串接在所述治疗性表位之后的相邻治疗性表位，并且P是由下式得到的路径矩阵：

$$\mathbf{P} = \begin{bmatrix} 0 & \mathbf{0}^{1\times v} \\ \mathbf{0}^{v\times 1} & \mathbf{D} \end{bmatrix},$$

其中D为v x v矩阵，其中元素D(k,m)指示所述一对排序的治疗性表位k,m的距离度量；以及

基于x_{km}的所述解出值来选择所述盒序列。

11.如权利要求1所述的方法，其还包括制造或已制造包含所述盒序列的肿瘤疫苗。

12.一种鉴别新抗原疫苗的盒序列的方法，其包括：

对于患者，从所述受试者的肿瘤细胞和正常细胞获得外显子组、转录组或全基因组肿瘤核苷酸测序数据中的至少一种，其中所述核苷酸测序数据被用于获得表示通过比较来自所述肿瘤细胞的核苷酸测序数据和来自所述正常细胞的核苷酸测序数据所鉴别的新抗原集合中每一者的肽序列的数据，其中每种新抗原的所述肽序列包含至少一个使其不同于从所述受试者的所述正常细胞鉴别的相应野生型亲本肽序列的变化，并且包含关于构成所述肽序列的多个氨基酸和所述肽序列中所述氨基酸的位置集合的信息；

对于所述受试者，鉴别来自所述新抗原集合的新抗原治疗子集；以及

对于所述受试者，鉴别包含串接的治疗性表位的序列的盒序列，所述治疗性表位各自包含所述新抗原治疗子集中相应新抗原的肽序列，其中所述盒序列基于跨越一对或多对相

邻治疗性表位之间的相应接合部的一个或多个接合表位的呈递来鉴别。

13.如权利要求12所述的方法,其中所述一个或多个接合表位的所述呈递基于通过将所述一个或多个接合表位的序列输入机器学习的呈递模型所生成的呈递可能性来确定,所述呈递可能性指示所述一个或多个接合表位由一个或多个MHC等位基因呈递在所述患者的肿瘤细胞表面上的可能性,所述呈递可能性集合已至少基于所接收的质谱数据来鉴别。

14.如权利要求12所述的方法,其中所述一个或多个接合表位的所述呈递基于所述受试者的所述一个或多个接合表位与一个或多个MHC等位基因之间的结合亲和力预测来确定。

15.如权利要求12所述的方法,其中所述一个或多个接合表位的所述呈递基于所述一个或多个接合表位的结合稳定性预测来确定。

16.如权利要求12所述的方法,其中所述一个或多个接合表位包括与第一治疗性表位的序列和串接在所述第一治疗性表位之后的第二治疗性表位的序列重叠的接合表位。

17.如权利要求12所述的方法,其中连接子序列置于第一治疗性表位与串接在所述第一治疗性表位之后的第二治疗性表位之间,并且所述一个或多个接合表位包括与所述连接子序列重叠的接合表位。

18.如权利要求12所述的方法,其中鉴别所述盒序列包括:

对于每对排序的治疗性表位,确定跨越所述一对排序的治疗性表位之间的所述接合部的接合表位集合;以及

对于每对排序的治疗性表位,确定指示所述排序对的所述接合表位集合在所述受试者的所述一个或多个MHC等位基因上的呈递的距离度量。

19.如权利要求12所述的方法,其中鉴别所述盒序列包括:

生成对应于所述治疗性表位的不同序列的候选盒序列集合;

对于每个候选盒序列,基于所述候选盒序列中每对排序的治疗性表位的所述距离度量确定所述候选盒序列的呈递评分;以及

选择与具有低于预定阈值的呈递评分相关联的候选盒序列为所述新抗原疫苗的所述盒序列。

20.如权利要求19所述的方法,其中随机生成所述候选盒序列集合。

21.如权利要求18所述的方法,其中鉴别所述盒序列还包括:

解出以下优化问题中x_{km}的值:

$$\underset{x}{\text{最小值}} \sum_{k=1}^{v+1} \sum_{k\neq m, m=1}^{v+1} P_{km} \cdot x_{km}$$

$$\sum_{k=1}^{v+1} x_{km} = 1, \quad m = 1, 2, \ldots, v+1$$

$$\sum_{m=1}^{v+1} x_{km} = 1, \quad k = 1, 2, \ldots, v+1$$

$x_{kk}=0, k=1,2,\ldots,v+1$

$$\text{out}(S) \geq 1, \quad S \subset E, 2 \leq |S| \leq |V|/2$$

其中v对应于新抗原的预定数目，k对应于治疗性表位并且m对应于串接在所述治疗性表位之后的相邻治疗性表位，并且P是由下式得到的路径矩阵：

$$\mathbf{P} = \begin{bmatrix} 0 & \mathbf{0}^{1\times v} \\ \mathbf{0}^{v\times 1} & \mathbf{D} \end{bmatrix},$$

其中D为v x v矩阵，其中元素D(k,m)指示所述一对排序的治疗性表位k,m的距离度量；以及

基于x_{km}的所述解出值来选择所述盒序列。

22.如权利要求12所述的方法，其还包括制造或已制造包含所述盒序列的肿瘤疫苗。

23.一种鉴别新抗原疫苗的盒序列的方法，其包括：

获得共有抗原的治疗子集或共有新抗原的治疗子集的肽序列用于治疗多个受试者，所述治疗子集对应于具有高于预定阈值的呈递可能性的肽序列的预定数目；以及

鉴别包含串接的治疗性表位的序列的所述盒序列，所述治疗性表位各自包含所述共有抗原治疗子集或所述共有新抗原治疗子集中的相应肽序列，其中鉴别所述盒序列包括：

对于每对排序的治疗性表位，确定跨越所述一对排序的治疗性表位之间的所述接合部的接合表位集合；以及

对于每对排序的治疗性表位，确定指示所述排序对的所述接合表位集合的呈递的距离度量，其中所述距离度量被确定为各自指示相应MHC等位基因的流行性的加权集合与指示所述接合表位集合在所述MHC等位基因上的呈递可能性的相应次距离度量的组合。

24.一种肿瘤疫苗，其包含含有串接的治疗性表位的序列的盒序列，所述盒序列通过执行以下步骤来鉴别：

对于患者，从所述受试者的肿瘤细胞和正常细胞获得外显子组、转录组或全基因组肿瘤核苷酸测序数据中的至少一种，其中所述核苷酸测序数据被用于获得表示通过比较来自所述肿瘤细胞的核苷酸测序数据和来自所述正常细胞的核苷酸测序数据所鉴别的新抗原集合中每一者的肽序列的数据，其中每种新抗原的所述肽序列包含至少一个使其不同于从所述受试者的所述正常细胞鉴别的相应野生型亲本肽序列的变化，并且包含关于构成所述肽序列的多个氨基酸和所述肽序列中所述氨基酸的位置集合的信息；

对于所述受试者，鉴别来自所述新抗原集合的新抗原治疗子集；以及

对于所述受试者，鉴别包含串接的治疗性表位的序列的盒序列，所述治疗性表位各自包含所述新抗原治疗子集中相应新抗原的肽序列，其中所述盒序列基于跨越一对或多对相邻治疗性表位之间的相应接合部的一个或多个接合表位的呈递来鉴别。

25.如权利要求24所述的肿瘤疫苗，其中所述一个或多个接合表位的所述呈递基于通过将所述一个或多个接合表位的序列输入机器学习的呈递模型所生成的呈递可能性来确定，所述呈递可能性指示所述一个或多个接合表位由一个或多个MHC等位基因呈递在所述患者的肿瘤细胞表面上的可能性，所述呈递可能性集合已至少基于所接收的质谱数据来鉴别。

26.如权利要求24所述的肿瘤疫苗，其中所述一个或多个接合表位的所述呈递基于所述受试者的所述一个或多个接合表位与一个或多个MHC等位基因之间的结合亲和力预测来确定。

27.如权利要求24所述的肿瘤疫苗，其中所述一个或多个接合表位的所述呈递基于所

述一个或多个接合表位的结合稳定性预测来确定。

28.如权利要求24所述的肿瘤疫苗,其中所述一个或多个接合表位包括与第一治疗性表位的序列和串接在所述第一治疗性表位之后的第二治疗性表位的序列重叠的接合表位。

29.如权利要求24所述的肿瘤疫苗,其中连接子序列置于第一治疗性表位与串接在所述第一治疗性表位之后的第二治疗性表位之间,并且所述一个或多个接合表位包括与所述连接子序列重叠的接合表位。

30.如权利要求24所述的肿瘤疫苗,其中鉴别所述盒序列包括:

对于每对排序的治疗性表位,确定跨越所述一对排序的治疗性表位之间的所述接合部的接合表位集合;以及

对于每对排序的治疗性表位,确定指示所述排序对的所述接合表位集合在所述受试者的所述一个或多个MHC等位基因上的呈递的距离度量。

31.如权利要求24所述的肿瘤疫苗,其中鉴别所述盒序列包括:

生成对应于所述治疗性表位的不同序列的候选盒序列集合;

对于每个候选盒序列,基于所述候选盒序列中每对排序的治疗性表位的所述距离度量确定所述候选盒序列的呈递评分;以及

选择与具有低于预定阈值的呈递评分相关联的候选盒序列为所述新抗原疫苗的所述盒序列。

32.如权利要求31所述的肿瘤疫苗,其中随机生成所述候选盒序列集合。

33.如权利要求30所述的肿瘤疫苗,其中鉴别所述盒序列还包括:

解出以下优化问题中x_{km}的值:

$$\underset{x}{\text{最小值}} \sum_{k=1}^{v+1} \sum_{k\neq m, m=1}^{v+1} P_{km} \cdot x_{km}$$

$$\sum_{k=1}^{v+1} x_{km} = 1, \quad m = 1, 2, \ldots, v+1$$

$$\sum_{m=1}^{v+1} x_{km} = 1, \quad k = 1, 2, \ldots, v+1$$

$x_{kk}=0, k=1,2,\ldots,v+1$

$$\text{out}(S) \geq 1, \quad S \subset E, 2 \leq |S| \leq |V|/2$$

其中v对应于新抗原的预定数目,k对应于治疗性表位并且m对应于串接在所述第一治疗性表位之后的相邻治疗性表位,并且P是由下式得到的路径矩阵:

$$\mathbf{P} = \begin{bmatrix} 0 & \mathbf{0}^{1\times v} \\ \mathbf{0}^{v\times 1} & \mathbf{D} \end{bmatrix},$$

其中D为v x v矩阵,其中元素D(k,m)指示所述一对排序的治疗性表位k,m的距离度量;以及

基于x_{km}的所述解出值来选择所述盒序列。

34.如权利要求24所述的肿瘤疫苗,其还包含制造或已制造包含盒序列的肿瘤疫苗。

35.一种肿瘤疫苗,其包含含有串接的治疗性表位的序列的盒序列,所述盒序列排序,

使得各自包含新抗原治疗子集的相应新抗原的肽序列,其中治疗性表位的序列基于跨越一对或多对相邻治疗性表位之间的相应接合部的一个或多个接合表位的呈递来鉴别,其中盒序列的接合表位具有低于阈值结合亲和力的HLA结合亲和力。

36.如权利要求35所述的肿瘤疫苗,其中所述阈值结合亲和力为1000nM或更大。

37.一种肿瘤疫苗,其包含含有串接的治疗性表位的序列的盒序列,所述盒序列排序,使得各自包含新抗原治疗子集的相应新抗原的肽序列,其中治疗性表位的序列基于跨越一对或多对相邻治疗性表位之间的相应接合部的一个或多个接合表位的呈递来鉴别,其中盒序列的接合表位的至少一个阈值百分比具有低于阈值呈递可能性的呈递可能性。

38.如权利要求37所述的肿瘤疫苗,其中所述阈值百分比为50%。

减少新抗原的接合表位呈递

[0001] 相关申请的交叉引用

[0002] 本申请要求2017年11月22日提交的美国临时申请62/590,045的权益和优先权,所述专利全文出于所有目的以引用方式并入。

背景技术

[0003] 基于肿瘤特异性新抗原的治疗性疫苗作为新一代个性化癌症免疫疗法具有广阔的前景。[1-3]鉴于产生新抗原的可能性相对较高,具有高突变负荷的癌症,如非小细胞肺癌(NSCLC)和黑素瘤成为此类疗法的特别值得关注的靶标。[4,5]早期有证据显示,基于新抗原的疫苗接种能够引起T细胞反应[6]并且靶向新抗原的细胞疗法在某些情况下能够在选择的患者中引起肿瘤消退。[7]I类MHC和II类MHC对T细胞反应具有影响[70-71]。

[0004] 新抗原疫苗设计的一个问题是在受试者肿瘤内存在的众多编码突变中,哪种突变可以产生“最佳的”治疗性新抗原,例如能够引起抗肿瘤免疫并使肿瘤消退的抗原。

[0005] 提出的初步方法并入了使用下一代测序的基于突变的分析、RNA基因表达和候选新抗原肽的MHC结合亲和力预测[8]。然而,提出的这些方法都无法模拟整个表位产生过程,所述过程除含有基因表达和MHC结合外,还含有许多步骤(例如TAP转运、蛋白酶体裂解、MHC结合、将肽-MHC复合物转运至细胞表面和/或TCR对MHC-I的识别;内吞或自噬、通过细胞外或溶酶体蛋白酶(例如组织蛋白酶)裂解、与CLIP肽竞争HLA-DM催化的HLA结合、将肽-MHC复合物转运至细胞表面和/或TCR对MHC-II的识别)[9]。因此,现有的方法可能会有低阳性预测值(PPV)降低的问题。(图1A)

[0006] 事实上,多个研究团队所进行的关于由肿瘤细胞呈递的肽的分析显示,预计使用基因表达和MHC结合亲和力呈递的肽中不到5%可以在肿瘤表面MHC上发现[10,11](图1B)。近期观察到的仅针对突变数量的检查点抑制剂反应无法提高对结合受限的新抗原的预测准确性进一步支持了结合预测与MHC呈递之间的这一低相关性。[12]

[0007] 现有的呈递预测方法的这一低阳性预测值(PPV)提出了有关基于新抗原的疫苗设计的问题。如果使用PPV低的预测方法来设计疫苗,则大多数患者不太可能接受治疗性新抗原,且少数患者可能要接受一种以上新抗原(即使假设所有呈递的肽都具有免疫原性)。因此,用当前方法进行新抗原疫苗接种不太可能在众多具有肿瘤的受试者中取得成功。(图1C)

[0008] 此外,先前的方法仅使用顺式作用突变来产生候选新抗原,而在很大程度上忽视了考虑neo-ORF的其它来源,包括在多种肿瘤类型中出现且导致许多基因异常剪接的剪接因子突变[13],以及产生或移除蛋白酶裂解位点的突变。

[0009] 由于文库构建、外显子组和转录组捕捉、测序或数据分析的条件并非最佳条件,故肿瘤基因组和转录组分析的标准方法可能会遗漏产生候选新抗原的体细胞突变。同样,标准肿瘤分析方法可能会无意中促成序列伪影或生殖系多态现象作为新抗原,而分别导致疫苗能力的低效使用或自身免疫的风险。

[0010] 新抗原疫苗通常也被设计为疫苗盒,其中一系列治疗性表位一个接一个地串接。

疫苗盒序列可包含或不包含相邻对的治疗性表位之间的连接子序列。盒序列可产生作为新型但不相关表位序列的接合表位序列的接合表位，其跨越一对治疗性表位之间的接合部。接合表位可能由患者的I类或II类HLA等位基因呈递，并分别刺激CD8或CD4 T细胞反应。此类反应经常是不希望的，因为与接合表位反应的T细胞不具有治疗益处，并且可通过抗原竞争减小对盒中选定的治疗性表位的免疫反应。

发明内容

[0011] 本文公开了一种鉴别和选择用于个性化癌症疫苗的新抗原的优化方法。首先，提出了使用下一代测序(NGS)鉴别新抗原候选物的优化的肿瘤外显子组和转录组分析方法。这些方法建立在标准NGS肿瘤分析方法的基础之上，以确保在所有类别的基因组变化内推进最高敏感性和特异性的新抗原候选物。其次，提出了选择高PPV新抗原的新颖方法来克服特异性问题并确保打算包括在疫苗中的新抗原较大可能地引发抗肿瘤免疫。取决于实施方案，这些方法包括训练的统计回归或非线性深度学习模型，这些模型共同地模拟肽-等位基因定位以及多种长度的肽的独立等位基因基元(per-allele motif)，在不同长度的肽中共有统计强度。非线性深度学习模型可以专门设计和训练用于将同一细胞中的不同MHC等位基因处理为独立的，由此解决了线性模型所具有的不同MHC等位基因会相互干扰的问题。最后，解决了基于新抗原的个性化疫苗设计和制造的其它需要考虑的问题。

[0012] 给定治疗性表位集合，设计盒序列来减小接合表位在患者中呈递的可能性。盒序列通过考虑跨越盒中一对治疗性表位之间的接合部的接合表位的呈递来设计。在一个实施方案中，盒序列基于各自与盒的接合部相关联的距离度量集合来设计。距离度量可指示将呈递跨越一对相邻表位之间的一个或多个接合表位的可能性。在一个实施方案中，一个或多个候选盒序列通过随机改变所述治疗性表位集合串接的顺序来生成，并且选择具有低于预定阈值的呈递评分(例如，距离度量总和)的盒序列。在另一个实施方案中，治疗性表位被建模为节点，并且一对相邻表位的距离度量表示相应节点之间的距离。选择导致仅“访问”每个治疗性表位一次的总距离低于预定阈值的盒序列。

附图说明

[0013] 参照以下描述和附图将更好地理解本发明的这些和其他特征、方面和优势，在附图中：

[0014] 图1A示出当前用于鉴别新抗原的临床方法。

[0015] 图1B示出<5％的预测结合肽被呈递在肿瘤细胞上。

[0016] 图1C示出新抗原预测特异性问题的影响。

[0017] 图1D示出结合预测不足以进行新抗原鉴别。

[0018] 图1E示出MHC-I呈递的机率随肽长度的变化。

[0019] 图1F示出由Promega动态范围标准(dynamic range standard)生成的示例性肽谱。

[0020] 图1G示出添加特征如何增加模型阳性预测值。

[0021] 图2A是根据一个实施方案，用于鉴别患者体内肽呈递的可能性的环境的概述。

[0022] 图2B和2C示出根据一个实施方案的获得呈递信息的方法。

[0023] 图3是示出根据一个实施方案的呈递鉴别系统的计算机逻辑组件的高级框图。

[0024] 图4示出根据一个实施方案的示例性训练数据集。

[0025] 图5示出与MHC等位基因相关联的示例性网络模型。

[0026] 图6A示出根据一个实施方案的MHC等位基因共有的示例性网络模型NNH(•)。图6B示出根据另一个实施方案的MHC等位基因共有的示例性网络模型NN_H(•)。

[0027] 图7示出使用示例性网络模型生成与一个MHC等位基因相关联的肽的呈递可能性。

[0028] 图8示出使用示例性网络模型生成与一个MHC等位基因相关联的肽的呈递可能性。

[0029] 图9示出使用示例性网络模型生成与多个MHC等位基因相关联的肽的呈递可能性。

[0030] 图10示出使用示例性网络模型生成与多个MHC等位基因相关联的肽的呈递可能性。

[0031] 图11示出使用示例性网络模型生成与多个MHC等位基因相关联的肽的呈递可能性。

[0032] 图12示出使用示例性网络模型生成与多个MHC等位基因相关联的肽的呈递可能性。

[0033] 图13示出确定两个示例性盒序列的距离度量。

[0034] 图14示出用于实施图1和3中所示的实体的示例性计算机。

具体实施方式

[0035] I.定义

[0036] 一般说来,权利要求书和说明书中使用的术语意图解释为具有与本领域普通技术人员所理解的普通含义。为清楚起见,以下定义某些术语。如果普通含义与所提供的定义之间存在矛盾,应使用所提供的定义。

[0037] 如本文所使用,术语“抗原”是诱导免疫反应的物质。

[0038] 如本文所使用,术语“新抗原”是具有至少一个使其不同于相应野生型亲本抗原的变化的抗原,例如,所述变化是肿瘤细胞突变或肿瘤细胞特异性翻译后修饰。新抗原可以包括多肽序列或核苷酸序列。突变可以包括移码或非移码插入缺失、错义或无义取代、剪接位点变化、基因组重排或基因融合,或产生neoORF的任何基因组或表达变化。突变还可以包括剪接变体。肿瘤细胞特异性翻译后修饰可以包括异常磷酸化。肿瘤细胞特异性翻译后修饰还可以包括蛋白酶体产生的剪接抗原。参见Liepe等人,A large fraction of HLA class I ligands are proteasome-generated spliced peptides;Science.2016 Oct 21;354(6310):354-358。

[0039] 如本文所使用,术语“肿瘤新抗原”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相应正常细胞或组织中的新抗原。

[0040] 如本文所使用,术语“基于新抗原的疫苗”是基于一个或多个新抗原,例如多个新抗原的疫苗构建体。

[0041] 如本文所使用,术语“候选新抗原”是产生可以代表新抗原的新序列的突变或其他异常。

[0042] 如本文所使用,术语“编码区”是基因中编码蛋白质的部分。

[0043] 如本文所使用,术语“编码突变”是在编码区中存在的突变。

[0044] 如本文所使用,术语“ORF”是指开放阅读框。

[0045] 如本文所使用,术语“NEO-ORF”是由突变或其他异常如剪接而产生的肿瘤特异性ORF。

[0046] 如本文所使用,术语“错义突变”是导致一个氨基酸被另一个氨基酸取代的突变。

[0047] 如本文所使用,术语“无义突变”是导致一个氨基酸被终止密码子取代的突变。

[0048] 如本文所使用,术语“移码突变”是导致蛋白质框架改变的突变。

[0049] 如本文所使用,术语“插入缺失”是一个或多个核酸的插入或缺失。

[0050] 如本文在两个或更多个核酸或多肽序列的情况下使用的术语“同一性”百分比是指当比较并对准达到最大对应性时,如使用以下描述的序列比较算法(例如BLASTP和BLASTN,或技术人员可用的其他算法)之一测量或通过目测检查得到的两个或更多个序列或子序列具有指定百分比的核苷酸或氨基酸残基是相同的。取决于应用,“同一性”百分比可以存在于所比较的序列的某一区域内,例如在功能结构域内,或者存在于待比较的两个序列的全长内。

[0051] 为进行序列比较,通常,一个序列充当参考序列,以与测试序列相比较。当使用序列比较算法时,将测试序列和参考序列输入计算机,必要时指定子序列座标,并且指定序列算法程序参数。序列比较算法随后基于指定的程序参数计算一个或多个测试序列相对于参考序列的百分比序列同一性。或者,可以通过组合在所选序列位置(例如序列基元)处特定核苷酸,或对于翻译的序列来说特定氨基酸的存在或不存在来确定序列相似性或不相似性。

[0052] 用于比较的序列的最佳比对可以例如通过以下方法进行:Smith和Waterman,Adv.Appl.Math.2:482(1981)的局部同源性算法;Needleman和Wunsch,J.Mol.Biol.48:443(1970)的同源性比对算法;Pearson和Lipman,Proc.Nat'l.Acad.Sci.USA 85:2444(1988)的对于相似性方法的探索;这些算法的计算机化实现(在Wisconsin Genetics Software Package中的GAP、BESTFIT、FASTA和TFASTA,Genetics Computer Group,575 Science Dr.,Madison,Wis.);或视觉检查(通常参见Ausubel等人,下文)。

[0053] 适于测定序列同一性和序列相似性百分比的算法的一个实例是Altschul等人,J.Mol.Biol.215:403-410(1990)中描述的BLAST算法。执行BLAST分析的软件通过National Center for Biotechnology Information公开可用。

[0054] 如本文所使用,术语“无终止或通读”是导致天然终止密码子移除的突变。

[0055] 如本文所使用,术语“表位”是抗原中通常由抗体或T细胞受体结合的特定部分。

[0056] 如本文所使用,术语“免疫原性”是例如通过T细胞、B细胞或两者引发免疫反应的能力。

[0057] 如本文所使用,术语“HLA结合亲和力”、“MHC结合亲和力”意思指特定抗原与特定MHC等位基因之间的结合亲和力。

[0058] 如本文所使用,术语“诱饵(bait)”是用于自样品富集特定DNA或RNA序列的核酸探针。

[0059] 如本文所使用,术语“变体”是受试者的核酸与用作对照的参考人基因组之间的差异。

[0060] 如本文所使用,术语“变体识别(variant call)”是对通常由测序确定的变体存在

的算法确定。

[0061] 如本文所使用,术语“多态现象”是生殖系变体,即,在个体的所有带有DNA的细胞中所发现的变体。

[0062] 如本文所使用,术语“体细胞变体”是在个体的非生殖系细胞中产生的变体。

[0063] 如本文所使用,术语“等位基因”是基因的一种形式,或是基因序列的一种形式,或是蛋白质的一种形式。

[0064] 如本文所使用,术语“HLA型”是HLA基因等位基因的互补序列。

[0065] 如本文所使用,术语“无义介导的衰变”或“NMD”是由过早终止密码子引起的细胞对mRNA的降解。

[0066] 如本文所使用,术语“躯干突变”是起源于肿瘤发展早期且存在于大多数肿瘤细胞中的突变。

[0067] 如本文所使用,术语“亚克隆突变”是起源于肿瘤发展后期且仅存在于一小部分肿瘤细胞中的突变。

[0068] 如本文所使用,术语“外显子组”是编码蛋白质的基因组的子组。外显子组可以是基因组的全体外显子。

[0069] 如本文所使用,术语“逻辑回归”是由统计得到的二进制数据的回归模型,其中因变量等于1的机率的分对数被建模为因变量的线性函数。

[0070] 如本文所使用,术语“神经网络”是用于分类或回归的机器学习模型,由多层线性变换,继之以通常通过随机梯度下降和反向传播训练的逐元素非线性组成。

[0071] 如本文所使用,术语“蛋白质组”是由细胞、细胞群或个体表达和/或翻译的所有蛋白质的集合。

[0072] 如本文所使用,术语“肽组”是由MHC-I或MHC-II呈递于细胞表面上的所有肽的集合。肽组可以指一个细胞或一组细胞(例如肿瘤肽组,意思指构成肿瘤的所有细胞的肽组的联合)的特性。

[0073] 如本文所使用,术语“ELISPOT”意思指酶联免疫吸附斑点测定,这是一种用于监测人和动物的免疫反应的常用方法。

[0074] 如本文所使用,术语“dextramer”是在流式细胞术中用于抗原特异性T细胞染色的基于葡聚糖的肽-MHC多聚体。

[0075] 如本文所使用,术语“耐受性或免疫耐受性”是对一种或多种抗原,例如自身抗原免疫无反应性的状态。

[0076] 如本文所使用,术语“中枢耐受性”是通过缺失自身反应性T细胞克隆或通过促进自身反应性T细胞克隆分化成免疫抑制性调控性T细胞(Treg)而在胸腺中经历的耐受性。

[0077] 如本文所使用,术语“外周耐受性”是通过使经历中枢耐受性而存活的自身反应性T细胞下调或无反应性(anergizing),或通过促进这些T细胞分化成Treg而在外周经历的耐受性。

[0078] 术语“样品”可以包括借助于包括静脉穿刺、排泄、射精、按摩、活组织检查、针抽吸、灌洗样品、刮取、手术切口或干预在内的手段,或本领域中已知的其他手段从受试者获取单个细胞或多个细胞,或细胞碎片,或体液等分试样。

[0079] 术语“受试者”涵盖细胞、组织或生物体、人或非人,无论是体内、离体还是体外,雄

性还是雌性的。术语受试者包括含人在内的哺乳动物。

[0080] 术语“哺乳动物”涵盖人和非人两种，并且包括但不限于人、非人灵长类动物、犬科动物、猫科动物、鼠科动物、牛科动物、马科动物和猪科动物。

[0081] 术语“临床因素”是指受试者状况，例如疾病活动性或严重程度的量度。“临床因素”涵盖受试者健康状况的所有标志物，包括非样品标志物，和/或受试者的其他特征，如但不限于年龄和性别。临床因素可以是能通过在确定条件下评价来自受试者的一个样品(或样品群)或受试者而获得的分数、一个值或一组值。临床因素也可以由标志物和/或如基因表达替代物之类其他参数进行预测。临床因素可以包括肿瘤类型、肿瘤亚型和吸烟史。

[0082] 缩写：MHC：主要组织相容性复合物；HLA：人白细胞抗原或人MHC基因座；NGS：下一代测序；PPV：阳性预测值；TSNA：肿瘤特异性新抗原；FFPE：福尔马林固定、石蜡包埋；NMD：无义介导的衰变；NSCLC：非小细胞肺癌；DC：树突状细胞。

[0083] 除非上下文另外清楚地规定，否则如本说明书和所附权利要求中所使用，单数形式“一个(种)(a/an)”和“所述”包括多个参照物。

[0084] 本文中未直接定义的任何术语应理解为具有与本发明领域内所理解的通常与之相关的含义。本文论述的某些术语是为了向从业人员描述本发明各方面的组合物、装置、方法等以及其制备或使用提供额外的指导。应了解，相同的事物可以按超过一种方式表示。因此，替代性措辞和同义词可以用于本文所论述的任一个或多个术语。无论本文中是否阐述或论述术语都无关紧要。提供了一些同义词或可取代的方法、材料等。除非明确陈述，否则对一个或数个同义词或等效物的叙述不排除其他同义词或等效物的使用。实例，包括术语实例的使用只是出于说明的目的，且并非在本文中限制本发明各方面的范围和含义。

[0085] 说明书正文内引用的所有参考文献、颁布的专利和专利申请都是以引用的方式整体并入本文中用于所有目的。

[0086] II.减少接合表位呈递的方法

[0087] 本文公开了用于鉴别新抗原疫苗的盒序列的方法。例如，一种此类方法可包括以下步骤：对于患者，从受试者的肿瘤细胞和正常细胞获得外显子组、转录组或全基因组肿瘤核苷酸测序数据中的至少一者，其中所述核苷酸测序数据用于获得表示通过比较来自肿瘤细胞的核苷酸测序数据与来自正常细胞的核苷酸测序数据来鉴别的新抗原集合中每一者的肽序列的数据，其中每个新抗原的所述肽序列包含至少一个使其不同于由受试者的正常细胞鉴别的相应野生型亲本肽序列的变化，并且包含关于构成肽序列的多个氨基酸和所述肽序列中所述氨基酸的位置集合的信息；使用计算机处理器将新抗原的肽序列输入机器学习的呈递模型中以生成所述新抗原集合的数字呈递可能性集合，在所述集合中每个呈递可能性表示相应新抗原由一个或多个MHC等位基因呈递在受试者的肿瘤细胞表面上的可能性。机器学习的呈递模型包括至少基于训练数据集鉴别的多个参数。所述训练数据集包括：对于样品集合中的每个样品，标记，其通过测量结合至被鉴别为存在于样品中的MHC等位基因集合中的至少一个MHC等位基因的肽的存在的质谱法来获得；对于每个所述样品，训练肽序列，其包含关于构成训练肽的多个氨基酸和所述训练肽序列中所述氨基酸的位置集合的信息；以及一个函数，所述函数表示作为输入接收的新抗原的肽序列与作为输出生产的呈递可能性之间的关系。所述方法可还包括以下步骤：对于受试者，鉴别来自所述新抗原集合的新抗原治疗子集，所述新抗原治疗子集对应于具有高于预定阈值的呈递可能性的预定数

目的新抗原;以及对于受试者,鉴别包含串接的治疗性表位的序列的盒序列,所述治疗性表位各自包含新抗原治疗子集中的相应新抗原的肽序列,其中盒序列基于跨越一对或多对相邻治疗性表位之间的相应接合部的一个或多个接合表位的呈递来鉴别。

[0088] 一个或多个接合表位的呈递可基于通过将一个或多个接合表位的序列输入机器学习的呈递模型中生成的呈递可能性来确定。

[0089] 一个或多个接合表位的呈递可基于受试者的一个或多个接合表位与一个或多个MHC等位基因之间的结合亲和力预测来确定。

[0090] 一个或多个接合表位的呈递可基于一个或多个接合表位的结合稳定性预测来确定。

[0091] 一个或多个接合表位可包括与第一治疗性表位的序列和串接在所述第一治疗性表位之后的第二治疗性表位的序列重叠的接合表位。

[0092] 连接子序列可置于第一治疗性表位与串接在所述第一治疗性表位之后的第二治疗性表位之间,并且一个或多个接合表位包括与连接子序列重叠的接合表位。

[0093] 鉴别盒序列可还包括以下步骤:对于每对排序的治疗性表位,确定跨越所述一对排序的治疗性表位之间的接合部的接合表位集合;以及对于每对排序的治疗性表位,确定指示所述排序对的接合表位集合在受试者的一个或多个MHC等位基因上的呈递的距离度量。

[0094] 鉴别盒序列可还包括以下步骤:生成对应于治疗性表位的不同序列的候选盒序列集合;对于每个候选盒序列,基于候选盒序列中每对排序的治疗性表位的距离度量来确定候选盒序列的呈递评分;以及选择与低于预定阈值的呈递评分相关联的候选盒序列为新抗原疫苗的盒序列。

[0095] 可随机生成所述候选盒序列集合。

[0096] 鉴别盒序列可还包括在以下优化问题中解出x_{km}的值的步骤:

[0097] $$\underset{x}{\text{最小值}} \sum_{k=1}^{v+1} \sum_{k\neq m, m=1}^{v+1} \mathrm{P}_{km} \cdot x_{km}$$

[0098] $$\sum_{k=1}^{v+1} x_{km} = 1, \quad m = 1, 2, \ldots, v+1$$

[0099] $$\sum_{m=1}^{v+1} x_{km} = 1, \quad k = 1, 2, \ldots, v+1$$

[0100] $x_{kk}=0,k=1,2,\cdots,v+1$

[0101] $$\mathrm{out}(S) \geq 1, \quad S \subset E, 2 \leq |S| \leq |V|/2$$

[0102] 其中v对应于新抗原的预定数目,k对应于治疗性表位并且m对应于串接在所述治疗性表位之后的相邻治疗性表位,并且P是由下式得到的路径矩阵:

[0103] $$\mathbf{P} = \begin{bmatrix} 0 & \mathbf{0}^{1\times v} \\ \mathbf{0}^{v\times 1} & \mathbf{D} \end{bmatrix},$$

[0104] 其中D为v x v矩阵,其中元素D(k,m)指示一对排序的治疗性表位k,m的距离度量;以及基于x_{km}的解出值来选择盒序列。

[0105] 所述方法可还包括制造或已制造包含盒序列的肿瘤疫苗的步骤。

[0106] 本文还公开了一种鉴别新抗原疫苗的盒序列的方法，所述方法包括以下步骤：对于患者，从受试者的肿瘤细胞和正常细胞获得外显子组、转录组或全基因组肿瘤核苷酸测序数据中的至少一者，其中所述核苷酸测序数据用于获得表示通过比较来自肿瘤细胞的核苷酸测序数据与来自正常细胞的核苷酸测序数据来鉴别的新抗原集合中每一者的肽序列的数据，其中每个新抗原的所述肽序列包含至少一个使其不同于由受试者的正常细胞鉴别的相应野生型亲本肽序列的变化，并且包含关于构成肽序列的多个氨基酸和所述肽序列中所述氨基酸的位置集合的信息；对于受试者，鉴别来自所述新抗原集合的新抗原治疗子集；以及对于受试者，鉴别包含串接的治疗性表位的序列的盒序列，每个治疗性表位包含新抗原治疗子集中的相应新抗原的肽序列，其中盒序列基于跨越一对或多对相邻治疗性表位之间的相应接合部的一个或多个接合表位的呈递来鉴别。

[0107] 一个或多个接合表位的呈递可基于通过将一个或多个接合表位的序列输入机器学习的呈递模型中所生成的呈递可能性来确定，所述呈递可能性指示一个或多个接合表位由一个或多个MHC等位基因呈递在患者的肿瘤细胞表面上的可能性，所述呈递可能性集合已至少基于所接收的质谱数据来鉴别。

[0108] 一个或多个接合表位的呈递可基于受试者的一个或多个接合表位与一个或多个MHC等位基因之间的结合亲和力预测来确定。

[0109] 一个或多个接合表位的呈递可基于一个或多个接合表位的结合稳定性预测来确定。

[0110] 一个或多个接合表位可包括与第一治疗性表位的序列和串接在所述第一治疗性表位之后的第二治疗性表位的序列重叠的接合表位。

[0111] 连接子序列可置于第一治疗性表位与串接在所述第一治疗性表位之后的第二治疗性表位之间，并且一个或多个接合表位包括与连接子序列重叠的接合表位。

[0112] 鉴别盒序列可还包括以下步骤：对于每对排序的治疗性表位，确定跨越所述一对排序的治疗性表位之间的接合部的接合表位集合；以及对于每对排序的治疗性表位，确定指示所述排序对的接合表位集合在受试者的一个或多个MHC等位基因上的呈递的距离度量。

[0113] 鉴别盒序列可还包括以下步骤：生成对应于治疗性表位的不同序列的候选盒序列集合；对于每个候选盒序列，基于候选盒序列中每对排序的治疗性表位的距离度量来确定候选盒序列的呈递评分；以及选择与低于预定阈值的呈递评分相关联的候选盒序列为新抗原疫苗的盒序列。

[0114] 可随机生成所述候选盒序列集合。

[0115] 鉴别盒序列可还包括在以下优化问题中解出xkm的值的步骤：

[0116] $$\underset{x}{\text{最小值}} \sum_{k=1}^{v+1} \sum_{k\neq m, m=1}^{v+1} \mathrm{P}_{km} \cdot x_{km}$$

[0117] $$\sum_{k=1}^{v+1} x_{km} = 1, \quad m = 1, 2, \ldots, v+1$$

[0118] $$\sum_{m=1}^{v+1} x_{km} = 1, \quad k = 1,2,\dots,v+1$$

[0119] $x_{kk}=0,k=1,2,\cdots,v+1$

[0120] $$\text{out}(S) \geq 1, \quad S \subset E, 2 \leq |S| \leq |V|/2$$

[0121] 其中v对应于新抗原的预定数目,k对应于治疗性表位并且m对应于串接在所述治疗性表位之后的相邻治疗性表位,并且P是由下式得到的路径矩阵:

[0122] $$\mathbf{P} = \begin{bmatrix} 0 & \mathbf{0}^{1\times v} \\ \mathbf{0}^{v\times 1} & \mathbf{D} \end{bmatrix},$$

[0123] 其中D为v x v矩阵,其中元素D(k,m)指示一对排序的治疗性表位k,m的距离度量;以及基于x_{km}的解出值来选择盒序列。

[0124] 所述方法可还包括已制造包含盒序列的肿瘤疫苗的步骤。

[0125] 本文还公开了一种鉴别新抗原疫苗的盒序列的方法,所述方法包括以下步骤:获得用于治疗多个受试者的共有抗原治疗子集或共有新抗原治疗子集的肽序列,所述治疗子集对应于具有高于预定阈值的呈递可能性的预定数目的肽序列;以及鉴别包含串接的治疗性表位的序列的盒序列,每个治疗性表位包含共有抗原治疗子集或共有新抗原治疗子集的相应肽序列,其中鉴别盒序列包括对于每对排序的治疗性表位确定跨越所述一对排序的治疗性表位之间的接合部的接合表位集合;以及对于每对排序的治疗性表位,确定指示所述排序对的所述接合表位集合的呈递的距离度量,其中所述距离度量被确定为各自指示相应MHC等位基因的流行性的加权集合与指示MHC等位基因上所述接合表位集合的呈递可能性的相应亚距离度量的组合。

[0126] 本文还公开了一种肿瘤疫苗,其包含含有串接的治疗性表位的序列的盒序列,所述盒序列通过执行以下步骤来鉴别:对于患者,从受试者的肿瘤细胞和正常细胞获得外显子组、转录组或全基因组肿瘤核苷酸测序数据中的至少一者,其中所述核苷酸测序数据用于获得表示通过比较来自肿瘤细胞的核苷酸测序数据与来自正常细胞的核苷酸测序数据来鉴别的新抗原集合中每一者的肽序列的数据,其中每个新抗原的肽序列包含至少一个使其不同于由受试者的正常细胞鉴别的相应野生型亲本肽序列的变化,并且包含关于构成肽序列的多个氨基酸和所述肽序列中所述氨基酸的位置集合的信息;对于受试者,鉴别来自所述新抗原集合的新抗原治疗子集;以及对于受试者,鉴别包含串接的治疗性表位的序列的盒序列,每个治疗性表位包含新抗原治疗子集中的相应新抗原的肽序列,其中盒序列基于跨越一对或多对相邻治疗性表位之间的相应接合部的一个或多个接合表位的呈递来鉴别。

[0127] 一个或多个接合表位的呈递基于通过将一个或多个接合表位的序列输入机器学习的呈递模型中所生成的呈递可能性来确定,所述呈递可能性指示一个或多个接合表位由一个或多个MHC等位基因呈递在患者的肿瘤细胞表面上的可能性,所述呈递可能性集合已至少基于所接收的质谱数据来鉴别。

[0128] 一个或多个接合表位的呈递可基于受试者的一个或多个接合表位与一个或多个MHC等位基因之间的结合亲和力预测来确定。

[0129] 一个或多个接合表位的呈递可基于一个或多个接合表位的结合稳定性预测来确

定。

[0130] 一个或多个接合表位可包括与第一治疗性表位的序列和串接在所述第一治疗性表位之后的第二治疗性表位的序列重叠的接合表位。

[0131] 连接子序列可置于第一治疗性表位与串接在所述第一治疗性表位之后的第二治疗性表位之间，并且一个或多个接合表位包括与连接子序列重叠的接合表位。

[0132] 鉴别盒序列可还包括以下步骤：对于每对排序的治疗性表位，确定跨越所述一对排序的治疗性表位之间的接合部的接合表位集合；以及对于每对排序的治疗性表位，确定指示所述排序对的接合表位集合在受试者的一个或多个MHC等位基因上的呈递的距离度量。

[0133] 鉴别盒序列可还包括以下步骤：生成对应于治疗性表位的不同序列的候选盒序列集合；对于每个候选盒序列，基于候选盒序列中每对排序的治疗性表位的距离度量来确定候选盒序列的呈递评分；以及选择与低于预定阈值的呈递评分相关联的候选盒序列为新抗原疫苗的盒序列。

[0134] 可随机生成所述候选盒序列集合。

[0135] 鉴别盒序列可还包括在以下优化问题中解出xkm的值的步骤：

[0136] $$\underset{x}{\text{最小值}}\sum_{k=1}^{v+1}\sum_{k\neq m,m=1}^{v+1}\mathrm{P}_{km}\cdot x_{km}$$

[0137] $$\sum_{k=1}^{v+1}x_{km}=1,\quad m=1,2,\dots,v+1$$

[0138] $$\sum_{m=1}^{v+1}x_{km}=1,\quad k=1,2,\dots,v+1$$

[0139] $x_{kk}=0,k=1,2,\cdots,v+1$

[0140] $$\mathrm{out}(S)\geq 1,\quad S\subset E,2\leq|S|\leq|V|/2$$

[0141] 其中v对应于新抗原的预定数目，k对应于治疗性表位并且m对应于串接在所述第一治疗性表位之后的相邻治疗性表位，并且P是由下式得到的路径矩阵：

[0142] $$\mathbf{P}=\begin{bmatrix}0 & \mathbf{0}^{1\times v}\\ \mathbf{0}^{v\times 1} & \mathbf{D}\end{bmatrix},$$

[0143] 其中D为v x v矩阵，其中元素D(k,m)指示一对排序的治疗性表位k,m的距离度量；以及基于x_{km}的解出值来选择盒序列。

[0144] 如权利要求24所述的肿瘤疫苗，还包含制造或已制造包含盒序列的肿瘤疫苗。

[0145] 本文还公开了一种肿瘤疫苗，其包含含有串接的治疗性表位的序列的盒序列，所述盒序列排序，使得各自包含新抗原治疗子集的相应新抗原的肽序列，其中治疗性表位的序列基于跨越一对或多对相邻治疗性表位之间的相应接合部的一个或多个接合表位的呈递来鉴别，其中盒序列的接合表位具有低于阈值结合亲和力的HLA结合亲和力。

[0146] 阈值结合亲和力可为1000NM或更大。

[0147] 本文还公开了一种肿瘤疫苗，其包含含有串接的治疗性表位的序列的盒序列，所述盒序列排序，使得各自包含新抗原治疗子集的相应新抗原的肽序列，其中治疗性表位的

序列基于跨越一对或多对相邻治疗性表位之间的相应接合部的一个或多个接合表位的呈递来鉴别,其中盒序列的接合表位的至少一个阈值百分比具有低于阈值呈递可能性的呈递可能性。

[0148] 阈值百分比可为50%。

[0149] III.鉴别新抗原中的肿瘤特异性突变

[0150] 本文还公开了用于鉴别某些突变(例如癌细胞中存在的变体或等位基因)的方法。确切地说,这些突变可以存在于患有癌症的受试者的癌细胞的基因组、转录组、蛋白质组或外显子组中,但不存在于受试者的正常组织中。

[0151] 如果肿瘤中的基因突变仅导致肿瘤中蛋白质的氨基酸序列改变,则认为这些突变可用于免疫靶向肿瘤。有用的突变包括:(1)导致蛋白质中的氨基酸不同的非同义突变;(2)通读突变,其中终止密码子被修饰或缺失,导致翻译得到在C末端具有新肿瘤特异性序列的较长蛋白质;(3)导致在成熟mRNA中包括内含子且由此产生独特肿瘤特异性蛋白质序列的剪接位点突变;(4)产生在2种蛋白质的接合处具有肿瘤特异性序列的嵌合蛋白的染色体重排(即,基因融合);(5)产生具有新肿瘤特异性蛋白质序列的新开放阅读框的移码突变或缺失。突变还可以包括非移码插入缺失、错义或无义取代、剪接位点变化、基因组重排或基因融合,或产生neoORF的任何基因组或表达变化中的一种或多种。

[0152] 在肿瘤细胞中具有突变的肽或由例如剪接位点突变、移码突变、通读突变或基因融合突变产生的突变多肽可以通过对肿瘤和正常细胞中的DNA、RNA或蛋白质进行测序来鉴别。

[0153] 突变还可以包括先前鉴别的肿瘤特异性突变。已知的肿瘤突变可以见于癌症体细胞突变目录(Catalogue of Somatic Mutations in Cancer,COSMIC)数据库。

[0154] 多种方法可用于检测个体的DNA或RNA中特定突变或等位基因的存在。本领域中的改进之处在于提供准确、容易且便宜的大规模SNP基因分型。举例来说,已描述若干技术,包括动态等位基因特异性杂交(DASH)、微板阵列对角线凝胶电泳(microplate array diagonal gel electrophoresis,MADGE)、焦磷酸测序、寡核苷酸特异性连接、TaqMan系统以及各种DNA“芯片”技术,如Affymetrix SNP芯片。这些方法通常通过PCR扩增靶基因区。一些其他的方法基于通过侵袭式裂解产生小信号分子,随后进行质谱法或固定化挂锁探针(padlock probe)和滚环扩增。本领域中已知用于检测特定突变的若干方法概述于下。

[0155] 基于PCR的检测手段可以包括同时多重扩增多个标志物。举例来说,本领域中众所周知,选择PCR引物产生尺寸不重叠且可以同时分析的PCR产物。或者,可用以不同方式标记且由此可以通过不同方式检测的引物扩增不同标志物。当然,基于杂交的检测手段能够以不同方式检测样品中的多个PCR产物。本领域中已知能够多重分析多个标志物的其他技术。

[0156] 已经开发出数种方法来促进基因组DNA或细胞RNA中单核苷酸多态性的分析。举例来说,可以通过使用专用的核酸外切酶抗性核苷酸检测单碱基多态性,如例如Mundy,C.R.(美国专利第4,656,127号)中所公开的。根据所述方法,与紧靠多态性位点3’端的等位基因序列互补的引物能够与从特定动物或人获得的靶分子杂交。如果靶分子上的多态性位点含有与存在的特定核酸外切酶抗性核苷酸衍生物互补的核苷酸,则该衍生物将被合并至杂交引物的末端上。此类合并使得引物对核酸外切酶具有抗性,并由此允许其检测。由于样品的核酸外切酶抗性衍生物的身份是已知的,故引物对核酸外切酶产生抗性的发现披露,靶分

子多态性位点中存在的核苷酸与反应中使用的核苷酸衍生物互补。该方法的优势在于，它不需要测定大量无关的序列数据。

[0157] 可以使用基于溶液的方法来确定多态性位点的核苷酸的身份。Cohen,D.等人(法国专利2,650,840;PCT申请第WO91/02087号)。如在美国专利第4,656,127号的Mundy方法中所述，采用与紧靠多态性位点3'端的等位基因序列互补的引物。所述方法使用标记过的双脱氧核苷酸衍生物来确定该位点的核苷酸的身份，如果与多态性位点的核苷酸互补，则所述核苷酸将被合并至引物末端上。Goelet,P.等人(PCT申请号92/15712)描述了称为基因位分析或GBA的替代性方法。Goelet,P.等人的方法使用了标记过的终止子和与在多态性位点3'端的序列互补的引物的混合物。由此通过存在于所评价靶分子的多态性位点中的核苷酸来确定合并的标记过的终止子并且所述终止子与存在于所评价靶分子的多态性位点中的核苷酸互补。与Cohen等人(法国专利2,650,840;PCT申请第WO91/02087号)的方法相比，Goelet,P.等人的方法可以是非均相测定，其中引物或靶分子被固定于固相。

[0158] 已描述数种引物引导的用于测定DNA中的多态性位点的核苷酸并入程序(Komher,J.S.等人,Nucl.Acids.Res.17:7779-7784(1989);Sokolov,B.P.,Nucl.Acids Res.18:3671(1990);Syvanen,A.-C.等人,Genomics 8:684-692(1990);Kuppuswamy,M.N.等人,Proc.Natl.Acad.Sci.(U.S.A.)88:1143-1147(1991);Prezant,T.R.等人,Hum.Mutat.1:159-164(1992);Ugozzoli,L.等人,GATA 9:107-112(1992);Nyren,P.等人,Anal.Biochem.208:171-175(1993))。这些方法与GBA的不同之处在于，它们利用并入经过标记的脱氧核苷酸来区别多态性位点处的碱基。在此类形式中，由于信号与并入的脱氧核苷酸的数量成比例，故在同一核苷酸的操作中出现的多态现象可以产生与所述操作的长度成比例的信号(Syvanen,A.-C.等人,Amer.J.Hum.Genet.52:46-59(1993))。

[0159] 许多方案直接从数百万个独立DNA或RNA分子中并行获得序列信息。实时单分子边合成边测序技术依赖于荧光核苷酸的检测，因为这些核苷酸被并入与测序模板互补的新生DNA链中。在一种方法中，将30-50个碱基长度的寡核苷酸以5'端共价锚定至玻璃盖玻片上。这些锚定链执行两种功能。首先，如果模板被配置成具有与表面结合的寡核苷酸互补的捕捉尾部，则其充当靶模板链的捕捉位点。这些锚定链还充当模板引导的引物延伸的引物，形成序列读取的基础。捕捉引物用作固定位点以便使用多个合成、检测以及染料-连接子化学裂解以移除染料的循环进行序列测定。每个循环由添加聚合酶/标记过得核苷酸混合物、冲洗、成像和染料裂解组成。在一种替代方法中，聚合酶被修饰成具有荧光供体分子并且被固定于玻璃载片上，而各核苷酸用附接至γ-磷酸的受体萤光部分进行颜色编码。当核苷酸被并入从头合成的链中时，所述系统检测荧光标记的聚合酶与荧光修饰的核苷酸之间的相互作用。还存在其他边合成边测序技术。

[0160] 任何适合的边合成边测序平台都可以用于鉴别突变。如上文所描述，目前有四个主要的边合成边测序平台：来自Roche/454 Life Sciences的基因组测序仪、来自Illumina/Solexa的1G分析仪、来自Applied BioSystems的SOLiD系统以及来自Helicos Biosciences的Heliscope系统。Pacific BioSciences和VisiGen Biotechnologies也描述过边合成边测序平台。在一些实施方案中，使所测序的多个核酸分子结合至支撑物(例如固体支撑物)上。为了将核酸固定于支撑物上，可以在模板的3'和/或5'端添加捕捉序列/通用引发位点。可以通过使捕捉序列与共价附接至支撑物的互补序列杂交而使核酸结合至支撑

物。捕捉序列(又称为通用捕捉序列)是与附接至支撑物的序列互补的核酸序列,所述序列还可以充当通用引物。

[0161] 作为捕捉序列的替代,可以将偶合对(如抗体/抗原、受体/配体,或抗生物素-生物素对,如例如美国专利申请第2006/0252077号中所述)的一个成员连接至各片段以将其捕捉在涂有该偶合对的相应第二成员的表面上。

[0162] 在捕捉后,可以例如实施例和美国专利第7,283,337号中所描述,通过例如单分子检测/测序,包括模板依赖性边合成边测序对所述序列进行分析。在边合成边测序时,使表面结合的分子在聚合酶存在下暴露于多个标记过得核苷酸三磷酸。模板序列由并入正在生长的链的3'端的标记过的核苷酸的顺序决定。这可以实时进行或者可以按分步重复模式进行。对于实时分析,可以将不同光学标记并入各核苷酸并且可以利用多种激光器刺激并入的核苷酸。

[0163] 测序还可以包括其他大规模平行测序或下一代测序(NGS)技术和平台。大规模平行测序技术和平台的其它实例有Illumina HiSeq或MiSeq、Thermo PGM或Proton、Pac Bio RS II或Sequel、Qiagen公司的Gene Reader和Oxford Nanopore MinION。可以使用当前其它类似的大规模平行测序技术,以及这些技术的改进形式。

[0164] 所有细胞类型或组织都可以用于获得用于本文所描述的方法中的核酸样品。举例来说,DNA或RNA样品可以从肿瘤或体液,例如利用已知技术(例如静脉穿刺)获得的血液,或唾液获得。或者,可以对干燥样品(例如毛发或皮肤)进行核酸测试。此外,可以从肿瘤获得一份测序样品,并且可以从正常组织获得另一份测序样品,其中正常组织与肿瘤同属相同组织类型。可以从肿瘤获得一份测序样品,并且可以从正常组织获得另一份测序样品,其中正常组织与肿瘤属于不同组织类型。

[0165] 肿瘤可以包括以下一种或多种:肺癌、黑素瘤、乳癌、卵巢癌、前列腺癌、肾癌、胃癌、结肠癌、睾丸癌、头颈癌、胰腺癌、脑癌、B细胞淋巴瘤、急性骨髓性白血病、慢性骨髓性白血病、慢性淋巴细胞性白血病和T细胞淋巴细胞性白血病、非小细胞肺癌和小细胞肺癌。

[0166] 或者,可以使用蛋白质质谱法鉴别或验证结合至肿瘤细胞上的MHC蛋白质的突变肽的存在。肽可以用酸从肿瘤细胞或从自肿瘤免疫沉淀的HLA分子洗脱,并且接着使用质谱法鉴别。

[0167] IV.新抗原

[0168] 新抗原可以包括核苷酸或多肽。举例来说,新抗原可以是编码多肽序列的RNA序列。因此,可用于疫苗中的新抗原包括核苷酸序列或多肽序列。

[0169] 本文公开了包含通过本文所公开的方法鉴别的肿瘤特异性突变的分离的肽、包含已知肿瘤特异性突变的肽,以及通过本文所公开的方法鉴别的突变多肽或其片段。新抗原肽可以在其编码序列背景下描述,其中新抗原包括编码相关多肽序列的核苷酸序列(例如DNA或RNA)。

[0170] 由新抗原核苷酸序列编码的一个或多个多肽可以包含以下至少一种:以低于1000nM的IC50值的与MHC的结合亲和力;对于长度是8-15个,即8、9、10、11、12、13、14或15个氨基酸的I类MHC肽,在所述肽内或附近存在促进蛋白酶体裂解的序列基元;以及存在促进TAP转运的序列基元。对于长度是6-30个,即6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29或30个氨基酸的II类MHC肽,在所述肽内或附近存在促进

通过细胞外或溶酶体蛋白酶(组织蛋白酶)的切割或HLA-DM催化的HLA结合的序列基元。

[0171] 一个或多个新抗原可以被呈递于肿瘤表面上。

[0172] 一个或多个新抗原可以在患肿瘤的受试者中具有免疫原性,例如能够在所述受试者体内引起T细胞反应或B细胞反应。

[0173] 在产生用于患肿瘤的受试者的疫苗的情况下,可以考虑排除在受试者体内诱导自体免疫反应的一个或多个新抗原。

[0174] 至少一个新抗原肽分子的尺寸可以包括但不限于约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、约16个、约17个、约18个、约19个、约20个、约21个、约22个、约23个、约24个、约25个、约26个、约27个、约28个、约29个、约30个、约31个、约32个、约33个、约34个、约35个、约36个、约37个、约38个、约39个、约40个、约41个、约42个、约43个、约44个、约45个、约46个、约47个、约48个、约49个、约50个、约60个、约70个、约80个、约90个、约100个、约110个、约120个或更多个氨基分子残基,以及由其中可衍生的任何范围。在特定实施例方案中,新抗原肽分子等于或少于50个氨基酸。

[0175] 新抗原肽和多肽可以:对于I类MHC是15个或更少残基长度并且通常由介于约8个与约11个之间的残基,特别是9个或10个残基组成;对于II类MHC是6-30个残基(包括端点在内)。

[0176] 必要时,可以通过若干方式设计出更长的肽。在一种情况下,当预测出或已知肽在HLA等位基因上呈递的可能性时,较长的肽可以由以下任一种组成:(1)朝各相应基因产物的N末端和C末端延伸2-5个氨基酸的个别呈递的肽;(2)一些或全部呈递肽与各自的延伸序列的串接。在另一情况下,当测序披露在肿瘤中存在较长的(>10个残基)新表位序列(例如由产生新颖肽序列的移码、通读或包括内含子引起)时,较长的肽将由以下组成:(3)由新颖肿瘤特异性氨基酸组成的整个延伸段,由此绕过了对基于计算或体外测试来选择HLA呈递最强的较短肽的需求。在两种情况下,较长链的使用使患者细胞能够进行内源性加工并且可以产生更有效的抗原呈递和T细胞反应的诱导作用。

[0177] 新抗原肽和多肽可以被呈递于HLA蛋白质上。在一些方面,新抗原肽和多肽是以高于野生型肽的亲和力呈递于HLA蛋白质上。在一些方面,新抗原肽或多肽的IC50值可以是至少低于5000nM、至少低于1000nM、至少低于500nM、至少低于250nM、至少低于200nM、至少低于150nM、至少低于100nM、至少低于50nM或更低。

[0178] 在一些方面,新抗原肽和多肽当施用给受试者时不会诱导自体免疫反应和/或激发免疫耐受性。

[0179] 还提供了包含至少两个或更多个新抗原肽的组合物。在一些实施方案中,所述组合物含有至少两个不同的肽。至少两个不同的肽可以来源于同一多肽。不同的多肽意味着,所述肽的长度、氨基酸序列或两者不同。这些肽来源于已知或被发现含有肿瘤特异性突变的任何多肽。可以作为新抗原肽的来源的适合多肽可以见于例如COSMIC数据库。COSMIC策划了有关人癌症中的体细胞突变的全面信息。肽含有肿瘤特异性突变。在一些方面,肿瘤特异性突变是特定癌症类型的驱动突变。

[0180] 具有所希望的活性或特性的新抗原肽和多肽可以被修饰成用于提供某些所希望的属性,例如改良的药理学特征,同时增加或至少保持未修饰肽的大体上所有生物活性以结合所希望的MHC分子并活化适当T细胞。举例来说,新抗原肽和多肽可以经历各种变化,如

保守性或非保守性取代,其中此类变化可能在其使用中提供某些优势,如改良的MHC结合、稳定性和呈递。保守性取代意思指氨基酸残基被在生物上和/或化学上类似的另一氨基酸残基置换,例如一个疏水性残基被另一个置换,或一个极性残基被另一个置换。取代包括如Gly、Ala;Val、Ile、Leu、Met;Asp、Glu;Asn、Gln;Ser、Thr;Lys、Arg;以及Phe、Tyr等的组合。单氨基酸取代的影响还可以使用D-氨基酸探测。此类修饰可以使用众所周知的肽合成程序进行,如例如Merrifield,Science 232:341-347(1986),Barany&Merrifield,The Peptides,Gross&Meienhofer编辑(N.Y.,Academic Press),第1-284页(1979);以及Stewart和Young,Solid Phase Peptide Synthesis,(Rockford,Ill.,Pierce),第2版(1984)中所述。

[0181] 用各种氨基酸模拟物或非天然氨基酸修饰肽和多肽特别适用于增加所述肽和多肽的体内稳定性。稳定性可以通过多种方式测定。举例来说,使用肽酶和各种生物介质如人血浆和血清测试稳定性。参见例如Verhoef等人,Eur.J.Drug Metab Pharmacokin.11:291-302(1986)。肽的半衰期可以使用25%人血清(v/v)测定,按常规方式测定。方案大致如下。在使用前,通过离心使汇集的人血清(AB型,未热灭活)脱脂。接着,用RPMI组织培养基将所述血清稀释至25%并用于测试肽稳定性。按预定时间间隔,取出少量反应溶液并添加至6%三氯乙酸水溶液或乙醇中。冷却混浊的反应样品(4℃),保持15分钟,然后离心以使沉淀的血清蛋白聚结。接着,通过反相HPLC,使用稳定性特异性色谱条件测定肽的存在。

[0182] 这些肽和多肽可以经过修饰以提供除改良的血清半衰期外的所希望的属性。举例来说,可以通过将这些肽连接至含有至少一个能够诱导T辅助细胞反应的表位的序列来增强其诱导CTL活性的能力。免疫原性肽/T辅助偶联物可以借助于间隔子分子连接。间隔子通常包含在生理条件下大体上不带电荷的相对较小的中性分子,如氨基酸或氨基酸模拟物。这些间隔子通常选自例如Ala、Gly或由非极性氨基酸或中性极性氨基酸组成的其他中性间隔子。应理解,任选存在的间隔子无需包含相同残基且因此可以是异低聚物或同低聚物。当存在时,间隔子通常是至少一个或二个残基,更通常是三个至六个残基。或者,可以在无间隔子情况下将肽连接至T辅助肽。

[0183] 新抗原肽可以直接地或通过间隔子在肽的氨基或羧基末端连接至T辅助细胞。新抗原肽或T辅助肽的氨基末端可以被酰基化。示例性T辅助肽包括破伤风类毒素830-843、流感307-319、疟疾环子孢子382-398和378-389。

[0184] 蛋白质或肽可以通过本领域技术人员已知的任何技术制备,包括通过标准分子生物学技术表达蛋白质、多肽或肽、从天然来源分离蛋白质或肽,或化学合成蛋白质或肽。先前已公开对应于各种基因的核苷酸和蛋白质、多肽和肽序列,并且可以见于本领域普通技术人员已知的计算机化数据库。一种此类数据库是位于美国国家卫生研究院(National Institutes of Health)网站的国家生物技术信息中心(National Center for Biotechnology Information)的Genbank和GenPept数据库。已知基因的编码区可以使用本文所公开或本领域普通技术人员已知的技术扩增和/或表达。或者,本领域技术人员已知蛋白质、多肽和肽的各种市售制剂。

[0185] 在另一方面,新抗原包括了编码新抗原肽或其部分的核酸(例如多核苷酸)。所述多核苷酸可以是例如单链和/或双链DNA、cDNA、PNA、CAN、RNA(例如mRNA),或多核苷酸的天然或稳定化形式,如例如具有硫代磷酸酯主链的多核苷酸,或其组合,并且所述多核苷酸可

以含有或可以不含内含子。又另一方面提供了一种能够表达多肽或其部分的表达载体。用于不同细胞类型的表达载体是本领域众所周知的并且可以在无过度实验情况下进行选择。一般来说，将DNA以适当取向和正确的表达阅读框插入表达载体，如质粒中。必要时，可以将DNA连接至能被所希望的宿主识别的适当转录和翻译调控性控制核苷酸序列，不过此类控制一般在表达载体中可用。接着，通过标准技术将载体插入宿主中。相关指导可见于例如Sambrook等人(1989)Molecular Cloning,A Laboratory Manual,Cold Spring Harbor Laboratory,Cold Spring Harbor,N.Y.

[0186] V.疫苗组合物

[0187] 本文还公开了一种能够引起特异性免疫反应，例如肿瘤特异性免疫反应的免疫原性组合物，例如疫苗组合物。疫苗组合物通常包含多个例如使用本文所描述的方法选择的新抗原。疫苗组合物又可以称为疫苗。

[0188] 疫苗可以含有个数在1个与30个之间的肽，即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或30个不同的肽；6、7、8、9、10 11、12、13或14个不同肽；或12、13或14个不同的肽。肽可以包括翻译后修饰。疫苗可以含有个数在1个与100个之间或更多个核苷酸序列，即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、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、100或更多个不同的核苷酸序列；6、7、8、9、10 11、12、13或14个不同的核苷酸序列；或12、13或14个不同的核苷酸序列。疫苗可以含有个数在1个与30个之间的新抗原序列，即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、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、100或更多个不同的新抗原序列；6、7、8、9、10 11、12、13或14个不同的新抗原序列；或12、13或14个不同的新抗原序列。

[0189] 在一个实施方案中，不同肽和/或多肽或编码其的核苷酸序列的选择使得这些肽和/或多肽能够与不同MHC分子，如不同的I类MHC分子和/或不同的II类MHC分子缔合。在一些方面，一种疫苗组合物包含能够与最常出现的I类MHC分子和/或II类MHC分子缔合的肽和/或多肽的编码序列。因此，疫苗组合物可以包含能够与至少2个优选的、至少3个优选的或至少4个优选的I类MHC分子和/或II类MHC分子缔合的不同片段。

[0190] 所述疫苗组合物能够引起特异性细胞毒性T细胞反应和/或特异性辅助T细胞反应。

[0191] 疫苗组合物还可以包含佐剂和/或载剂。有用的佐剂和载剂的实例提供于下文中。组合物可以与载剂缔合，如例如蛋白质或抗原呈递细胞，如能够将肽呈递至T细胞的树突状细胞(DC)。

[0192] 佐剂是混合至疫苗组合物中增加或以其他方式改变针对新抗原的免疫反应的任何物质。载剂可以是能够与新抗原缔合的支架结构，例如多肽或多糖。任选地，佐剂是共价或非共价缀合的。

[0193] 佐剂增加针对抗原的免疫反应的能力通常通过免疫介导的反应的显著或实质上增加,或疾病症状的减少来表现。举例来说,体液免疫的增加通常表现为针对抗原所产生的抗体的效价的显著增加,并且T细胞活性增加通常表现为细胞增殖,或细胞毒性,或细胞因子分泌的增加。佐剂也可以通过例如将主要体液或Th反应变成主要细胞或Th反应来改变免疫反应。

[0194] 适合的佐剂包括但不限于,1018ISS、矾、铝盐、Amplivax、AS15、BCG、CP-870,893、CpG7909、CyaA、dSLIM、GM-CSF、IC30、IC31、咪喹莫特(Imiquimod)、ImuFact IMP321、IS Patch、ISS、ISCOMATRIX、JuvImmune、LipoVac、MF59、单磷酰脂质A、Montanide IMS 1312、Montanide ISA 206、Montanide ISA 50V、Montanide ISA-51、OK-432、OM-174、OM-197-MP-EC、ONTAK、PepTel载体系统、PLG微粒、雷西莫特(resiquimod)、SRL172、病毒颗粒和其他类病毒颗粒、YF-17D、VEGF捕捉剂、R848、β-葡聚糖、Pam3Cys、Aquila的来源于皂素的QS21刺激子(Aquila Biotech,Worcester,Mass.,USA)、分枝杆菌提取物和合成细菌细胞壁模拟物,以及其他专用佐剂,如Ribi的Detox.Quil或Superfos。佐剂,如不完全弗氏佐剂或GM-CSF是有用的。先前已描述若干专用于树突状细胞的免疫佐剂(例如MF59)和其制备方法(Dupuis M等人,Cell Immunol.1998;186(1):18-27;Allison A C;Dev Biol Stand.1998;92:3-11)。也可以使用细胞因子。若干细胞因子与以下直接相关:影响树突状细胞向淋巴组织(例如TNF-α)的迁移;加速树突状细胞成熟成为T淋巴细胞的有效抗原呈递细胞(例如GM-CSF、IL-1和IL-4)(美国专利第5,849,589号,特定地以引用的方式整体并入本文中)并充当免疫佐剂(例如IL-12)(Gabrilovich D I等人,J Immunother Emphasis Tumor Immunol.1996(6):414-418)。

[0195] 也已经报导过CpG免疫刺激性寡核苷酸能增强佐剂在疫苗环境中的作用。也可以使用其他TLR结合分子,如RNA结合性TLR 7、TLR 8和/或TLR 9。

[0196] 有用佐剂的其他实例包括但不限于,化学修饰的CpG(例如CpR、Idera)、聚(I:C)(例如聚i:CI2U)、非CpG细菌DNA或RNA以及免疫活性小分子和抗体,如环磷酰胺、舒尼替尼(sunitinib)、贝伐单抗(bevacizumab)、西乐葆(celebrex)、NCX-4016、西地那非(sildenafil)、他达那非(tadalafil)、伐地那非(vardenafil)、索拉非尼(sorafinib)、XL-999、CP-547632、帕佐盘尼(pazopanib)、ZD2171、AZD2171、伊匹单抗(ipilimumab)、曲美单抗(tremelimumab)和SC58175,这些可以起到治疗作用和/或充当佐剂。佐剂和添加剂的量和浓度可以由熟练技术人员容易地确定,无需过度实验。其它佐剂包括集落刺激因子,如粒细胞巨噬细胞集落刺激因子(GM-CSF,沙格司亭(sargramostim))。

[0197] 疫苗组合物可以包含超过一种不同的佐剂。此外,治疗组合物可以包含任何佐剂物质,包括上述任一种或其组合。另外,预期疫苗和佐剂可以一起施用或按任何适当的次序分开施用。

[0198] 载剂(或赋形剂)可以独立于佐剂而存在。载剂的功能可以是例如增加特定突变体的分子量以增加活性或免疫原性;赋予稳定性、增加生物活性或增加血清半衰期。此外,载剂可以帮助将肽呈递至T细胞。载剂可以是本领域技术人员已知的任何适合的载剂,例如蛋白质或抗原呈递细胞。载剂蛋白可以是但不限于匙孔血蓝蛋白、血清蛋白如转铁蛋白、牛血清白蛋白、人血清白蛋白、甲状腺球蛋白或卵白蛋白、免疫球蛋白或激素,如胰岛素或棕榈酸。对于人的免疫,载剂一般是对人生理学上可接受的载剂并且是安全的。不过,破伤风类

毒素和/或白喉类毒素是适合的载剂。或者，载剂可以是葡聚糖，例如琼脂糖。

[0199] 细胞毒性T细胞(CTL)识别呈结合至MHC分子的肽形式的抗原，而非整个外来抗原本身。MHC分子本身位于抗原呈递细胞的细胞表面上。因此，如果存在肽抗原、MHC分子和APC的三聚体复合物，则可能活化CTL。相应地，如果所述肽不仅用于活化CTL，而且如果另外添加具有相应MHC分子的APC，则其可以增强免疫反应。因此，在一些实施方案中，疫苗组合物另外含有至少一种抗原呈递细胞。

[0200] 新抗原也可以被包括在基于病毒载体的疫苗平台中，如牛痘、禽痘、自复制型α病毒、马拉巴病毒(marabavirus)、腺病毒(参见例如Tatsis等人，Adenoviruses，Molecular Therapy(2004)10，616—629)或慢病毒，包括但不限于第二代、第三代和/或混合第二/第三代慢病毒和设计成靶向特定细胞类型或受体的任何一代重组慢病毒(参见例如，Hu等人，Immunization Delivered by Lentiviral Vectors for Cancer and Infectious Diseases，Immunol Rev.(2011)239(1):45-61；Sakuma等人，Lentiviral vectors:basic to translational，Biochem J.(2012)443(3):603-18；Cooper等人，Rescue of splicing-mediated intron loss maximizes expression in lentiviral vectors containing the human ubiquitin C promoter，Nucl.Acids Res.(2015)43(1):682-690；Zufferey等人，Self-Inactivating Lentivirus Vector for Safe and Efficient In Vivo Gene Delivery，J.Virol.(1998)72(12):9873-9880)。取决于以上提到的基于病毒载体的疫苗平台的包装能力，此方法可以递送编码一个或多个新抗原肽的一个或多个核苷酸序列。这些序列可以侧接非突变序列，可以由连接子分开，或者可以在前面具有一个或多个靶向亚细胞区室的序列(参见例如，Gros等人，Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients，Nat Med.(2016)22(4):433-8；Stronen等人，Targeting of cancer neoantigens with donor-derived T cell receptor repertoires，Science.(2016)352(6291):1337-41；Lu等人，Efficient identification of mutated cancer antigens recognized by T cells associated with durable tumor regressions，Clin Cancer Res.(2014)20(13):3401-10)。在引入宿主中后，受感染的细胞表达新抗原，并由此引起针对肽的宿主免疫(例如CTL)反应。可用于免疫方案的牛痘载体和方法描述于例如美国专利第4，722，848号中。另一载体是卡介苗(Bacille Calmette Guerin，BCG)。BCG载体描述于Stover等人(Nature 351:456-460(1991))中。根据本文的描述，本领域技术人员将显而易见可用于新抗原的治疗性施用或免疫的多种其他疫苗载体，例如，伤寒沙门氏菌(Salmonella typhi)载体。

[0201] V.A.新抗原盒

[0202] 用于选择一个或多个新抗原、克隆并构建“盒”以及将其插入病毒载体的方法为本领域中给予本文所提供的教导内容的技术人员的技能内。“新抗原盒”意指一个或多个所选新抗原与转录所述一个或多个新抗原并表达转录产物所必需的其他调控元件的组合。一个或多个新抗原可以允许转录的方式可操作地连接至调控元件。此类元件包括可驱动一个或多个新抗原在用病毒载体转染的细胞中表达的常规调控元件。因此，新抗原盒也可含有连接至所述一个或多个新抗原并与其他任选调控元件一起位于重组载体的所选病毒序列内的所选启动子。

[0203] 有用启动子可为组成型启动子或调控型(诱导型)启动子，其将能够控制有待表达

的一个或多个新抗原的量。例如,期望的启动子是巨细胞病毒立即早期基因启动子/增强子[参见例如Boshart等人,Cell,41:521-530(1985)]。另一种期望的启动子包括劳斯氏肉瘤病毒LTR启动子/增强子。又一种期望的启动子/增强子序列未鸡细胞质β肌动蛋白启动子[T.A.Kost等人,Nucl.Acids Res.,11(23):8287(1983)]。其他适合或期望的启动子可由本领域的技术人员选择。

[0204] 新抗原盒还可包括与病毒载体序列异源的核酸序列,包括提供使转录物有效聚腺苷酸化(poly-A或pA)的信号的序列和具有功能性剪接供体和受体位点的内含子。用于本发明的示例性载体中的普通poly-A序列为来源于乳多空病毒SV-40的序列。poly-A序列通常可插入在盒内的基于新抗原的序列之后和病毒载体序列之前。普通内含子序列也可以来源SV-40,并且被称为SV-40T内含子序列。新抗原盒还可含有位于启动子/增强子序列与一个或多个新抗原之间的此内含子。这些和其他普通载体元件的选择是常规的[参见例如Sambrook等人,"Molecular Cloning.A Laboratory Manual.",第2版,Cold Spring Harbor Laboratory,New York(1989)以及其中引用的参考文献],并且许多此类序列可获自商业和工业来源以及获自Genbank。

[0205] 新抗原盒可具有一个或多个新抗原。例如,给定盒可包含1-10、1-20、1-30、10-20、15-25、15-20、1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20或更多个新抗原。新抗原可直接彼此连接。新抗原还可通过连接子彼此连接。新抗原可呈相对于彼此的任何取向,包括N至C或C至N。

[0206] 如上文所述,新抗原盒可位于病毒载体的任何所选缺失位点,诸如E1基因区域缺失或E3基因区域缺失位点以及可选择的其他位点。

[0207] V.B.免疫检查点

[0208] 本文所述的载体诸如本文所述的C68载体或本文所述的甲病毒载体可包含编码至少一个新抗原的核酸并且相同或单独的载体可包含编码至少一种结合至免疫检查点分子并阻断其活性的免疫调节剂(例如抗体,诸如scFv)的核酸。载体可包含新抗原盒和编码检查点抑制剂的一个或多个核酸分子。

[0209] 可靶向阻断或抑制的说明性免疫检查点分子包括但不限于CTLA-4、4-1BB(CD137)、4-1BBL(CD137L)、PDL1、PDL2、PD1、B7-H3、B7-H4、BTLA、HVEM、TIM3、GAL9、LAG3、TIM3、B7H3、B7H4、VISTA、KIR、2B4(属于CD2分子家族并且在所有NK、γδ和记忆CD8+(αβ)T细胞上表达)、CD160(也称为BY55)和CGEN-15049。免疫检查点抑制剂包括结合至以下一种或多种并阻断或抑制其活性的抗体或其抗原结合片段或其他结合蛋白:CTLA-4、PDL1、PDL2、PD1、B7-H3、B7-H4、BTLA、HVEM、TIM3、GAL9、LAG3、TIM3、B7H3、B7H4、VISTA、KIR、2B4、CD160和CGEN-15049。说明性免疫检查点抑制剂包括曲美单抗(Tremelimumab)(CTLA-4阻断抗体)、抗OX40、PD-L1单克隆抗体(抗B7-H1;MEDI4736)、伊匹单抗(ipilimumab)、MK-3475(PD-1阻断剂)、纳武单抗(Nivolumamb)(抗PD1抗体)、CT-011(抗PD1抗体)、BY55单克隆抗体、AMP224(抗PDL1抗体)、BMS-936559(抗PDL1抗体)、MPLDL3280A(抗PDL1抗体)、MSB0010718C(抗PDL1抗体)和易普利姆玛/伊匹单抗(Yervoy/ipilimumab)(抗CTLA-4检查点抑制剂)。抗体编码序列可使用本领域普通技术工程化到载体诸如C68中。示例性方法描述于Fang等人,Stable antibody expression at therapeutic levels using the 2A peptide.Nat Biotechnol.2005年5月;23(5):584-90.电子版2005年4月17日;所述文献出于所有目的以

引用方式并入本文。

[0210] V.A.有关疫苗设计和制造的其它考虑因素

[0211] V.A.1.确定涵盖所有肿瘤亚克隆的肽集合

[0212] 躯干肽(Truncal peptide),意思指由所有或大部分肿瘤亚克隆呈递的肽,将优先被包括在疫苗中。[53]任选地,如果不存在预测会以较高机率呈递并具有免疫原性的躯干肽,或者如果预测能够以较高机率呈递并具有免疫原性的躯干肽的数量足够小以致可以在疫苗中包括其它非躯干肽,则可以通过估计肿瘤亚克隆的数量和属性并选择肽以使所述疫苗所涵盖的肿瘤亚克隆的数量最大来对其它肽进行优先排序。[54]

[0213] V.A.2.新抗原优先排序

[0214] 与疫苗技术可以支持的量相比,在应用所有以上新抗原过滤器后,仍有许多候选新抗原可包括在疫苗中。另外,可以保留有关新抗原分析的各个方面的不确定性,并且在候选疫苗新抗原的不同特性之间可能存在折中。因此,可以考虑用整合式多维模型代替在选择过程的每个步骤中的预定过滤器,所述多维模型将候选新抗原放入具有至少以下轴的空间中并使用整合方法优化选择。

[0215] 1.自体免疫或耐受的风险(生殖系的风险)(通常优选较低的自体免疫风险)

[0216] 2.测序伪影的机率(通常优选较低的伪影机率)

[0217] 3.免疫原性的机率(通常优选较高的免疫原性机率)

[0218] 4.呈递机率(通常优选较高的呈递机率)

[0219] 5.基因表达(通常优选较高表达)

[0220] 6.HLA基因的覆盖率(参与呈递新抗原集合的HLA分子数量增多可以降低肿瘤通过HLA分子下调或突变而逃避免疫攻击的机率)

[0221] 7.HLA类别的覆盖率(同时覆盖HLA-I和HLA-II可能会增加治疗反应的几率并降低肿瘤逃逸的几率)

[0222] VI.治疗和制造方法

[0223] 还提供了一种通过向受试者施用一个或多个新抗原,如使用本文所公开的方法鉴别的多个新抗原来诱导受试者的肿瘤特异性免疫反应、针对肿瘤接种疫苗、治疗和或缓解受试者的癌症症状的方法。

[0224] 在一些方面,受试者被诊断患有癌症或有发生癌症的风险。受试者可以是需要肿瘤特异性免疫反应的人、狗、猫、马或任何动物。肿瘤可以是任何实体肿瘤,如乳房肿瘤、卵巢肿瘤、前列腺肿瘤、肺肿瘤、肾肿瘤、胃肿瘤、结肠肿瘤、睾丸肿瘤、头颈部肿瘤、胰腺肿瘤、脑肿瘤、黑素瘤和其他组织器官肿瘤;以及血液肿瘤,如淋巴瘤和白血病,包括急性骨髓性白血病、慢性骨髓性白血病、慢性淋巴细胞性白血病、T细胞淋巴细胞性白血病和B细胞淋巴瘤。

[0225] 新抗原的施用量应足以诱导CTL反应。

[0226] 新抗原可以单独施用或与其他治疗剂组合施用。治疗剂是例如化学治疗剂、放射或免疫疗法。针对特定癌症的任何适合的治疗性治疗都可以施用。

[0227] 此外,还可以向受试者施用抗免疫抑制/免疫刺激剂,如检查点抑制剂。举例来说,还可以向受试者施用抗CTLA抗体或抗PD-1或抗PD-L1。抗体阻断CTLA-4或PD-L1可以增强针对患者体内癌细胞的免疫反应。确切地说,经显示,当遵循疫苗接种方案时,有效阻断CTLA-

4。

[0228] 可以确定包括在疫苗组合物中的各新抗原的最佳量和最佳剂量方案。举例来说，可以制备供静脉内(i.v.)注射、皮下(s.c.)注射、皮内(i.d.)注射、腹膜内(i.p.)注射、肌肉内(i.m.)注射的新抗原或其变体。注射方法包括皮下(s.c.)、皮内(i.d.)、腹腔(i.p.)、肌内(i.m.)和静脉内。DNA或RNA注射方法包括皮内、肌内、皮下、腹腔和静脉内。本领域技术人员已知施用疫苗组合物的其他方法。

[0229] 疫苗可以被设计成使得组合物中存在的新抗原的选择、数量和/或量具有组织、癌症和/或患者特异性。举例来说，肽的确切选择可以由给定组织中亲本蛋白质的表达模式来指导。所述选择可以取决于癌症的具体类型、疾病状态、先前的治疗方案、患者的免疫状态和当然要考虑的患者的HLA单倍型。此外，根据特定患者的个人需要，疫苗还可以含有个性化组分。实例包括根据特定患者体内新抗原的表达来改变新抗原的选择或遵循第一轮治疗方案调整后续治疗。

[0230] 对于打算用作癌症疫苗的组合物，在正常组织中大量表达的具有类似正常自身肽的新抗原应当避免或以少量存在于本文所描述的组合物中。另一方面，如果已知患者的肿瘤大量表达某一新抗原，则用于治疗此癌症的相应药物组合物可以大量存在和/或可以包括超过一种对于此特定新抗原或此新抗原的路径具有特异性的新抗原。

[0231] 可以将包含新抗原的组合物施用给患上癌症的个体。在治疗应用中，组合物是以足以引起针对肿瘤抗原的有效CTL反应和治愈或至少部分停滞症状和/或并发症的量施用给患者。适于实现此目的的量定义为“治疗有效剂量”。有效用于此用途的量将取决于例如组成、施用方式、所治疗的疾病的分期和严重程度、患者的体重和一般健康状态，以及处方医师的判断。应了解，组合物一般可以用于严重疾病状态，也就是说，危及生命或可能危及生命的状况，特别是当癌症已经转移的时候。在此类情况下，考虑到要使外来物质最少以及新抗原的相对无毒性质，治疗医师有可能并且会感觉需要施用大体上过量的这些组合物。

[0232] 对于治疗用途，施用可以在检测到或手术移除肿瘤时开始。这之后是增加剂量，直到至少症状大体上减轻并且之后持续一段时间。

[0233] 用于治疗性治疗的药物组合物(例如疫苗组合物)意图用于肠胃外、表面、鼻、口或局部施用。药物组合物可以通过肠胃外施用，例如静脉内、皮下、皮内或肌肉内施用。这些组合物可以施用到手术切除的部位处以诱导针对肿瘤的局部免疫反应。本文公开了供肠胃外施用的组合物，这些组合物包含新抗原溶液并且疫苗组合物被溶解或悬浮于可接受的载剂，例如水性载剂中。可以使用多种水性载剂，例如水、缓冲水、0.9％生理盐水、0.3％甘氨酸、透明质酸等。这些组合物可以通过众所周知的常规灭菌技术灭菌，或者可以经历无菌过滤。由此得到的水溶液可以被包装起来按原样使用，或者被冻干；冻干的制剂在施用之前与无菌溶液组合。必要时，这些组合物可以含有药学上可接受的辅助物质以接近生理条件，如pH调节剂和缓冲剂、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。

[0234] 新抗原还可以通过脂质体施用，使脂质体靶向特定细胞组织，如淋巴组织。脂质体还可用于增加半衰期。脂质体包括乳液、泡沫状物、胶束、不溶性单层、液晶、磷脂分散体、薄层状层等。在这些制剂中，待递送的新抗原是单独或与结合至例如淋巴细胞间普遍存在的受体的分子如结合至CD45抗原的单克隆抗体，或与其他治疗或免疫原性组合物缀合作为脂

质体的一部分并入。因此，填充有所希望的新抗原的脂质体可以被引导至淋巴细胞部位，接着脂质体递送所选治疗性/免疫原性组合物。脂质体可以由标准囊泡形成脂质形成，这些脂质一般包括中性和带负电的磷脂以及固醇如胆固醇。脂质的选择一般通过考虑例如脂质体尺寸、酸不稳定性和脂质体在血流中的稳定性来指导。如例如Szoka等人，Ann.Rev.Biophys.Bioeng.9;467(1980)；美国专利第4,235,871号、第4,501,728号、第4,501,728号、第4,837,028号和第5,019,369号中所述，有多种可用于制备脂质体的方法。

[0235] 为靶向免疫细胞，打算并入脂质体中的配体可以包括例如对所希望的免疫系统细胞的细胞表面决定子具有特异性的抗体或其片段。脂质体悬浮液可以经静脉内、局部、表面等途径施用，其剂量尤其根据施用方式、所递送的肽和所治疗疾病的分期等而变化。

[0236] 出于治疗或免疫接种目的，还可以向患者施用编码肽的核酸和任选地一种或多种本文所描述的肽。常常使用多种方法将核酸递送给患者。举例来说，核酸可以直接被递送，如“裸DNA”。这一方法描述于例如Wolff等人,Science 247:1465-1468(1990)以及美国专利第5,580,859号和第5,589,466号。核酸还可以使用例如美国专利第5,204,253号中所描述的冲击递送法(ballistic delivery)施用。可以施用仅包含DNA的粒子。或者，可以使DNA附着至粒子，如金粒子。用于递送核酸序列的方法可以包括病毒载体、mRNA载体和DNA载体，利用或不利用电穿孔。

[0237] 核酸也可以与阳离子性化合物，如阳离子性脂质形成复合物来递送。脂质介导的基因递送方法描述于例如9618372WOAWO 96/18372;9324640WOAWO 93/24640;Mannino和Gould-Fogerite,BioTechniques 6(7):682-691(1988)；美国专利第5,279,833号;Rose美国专利第5,279,833号；9106309WOAWO 91/06309；以及Felgner等人，Proc.Natl.Acad.Sci.USA 84:7413-7414(1987)。

[0238] 新抗原也可以被包括在基于病毒载体的疫苗平台中，如牛痘、禽痘、自复制型α病毒、马拉巴病毒(marabavirus)、腺病毒(参见例如Tatsis等人,Adenoviruses,Molecular Therapy(2004)10,616—629)或慢病毒，包括但不限于第二代、第三代和/或混合第二/第三代慢病毒和设计成靶向特定细胞类型或受体的任何一代重组慢病毒(参见例如，Hu等人，Immunization Delivered by Lentiviral Vectors for Cancer and Infectious Diseases,Immunol Rev.(2011)239(1):45-61;Sakuma等人,Lentiviral vectors:basic to translational,Biochem J.(2012)443(3):603-18;Cooper等人,Rescue of splicing-mediated intron loss maximizes expression in lentiviral vectors containing the human ubiquitin C promoter,Nucl.Acids Res.(2015)43(1):682-690;Zufferey等人,Self-Inactivating Lentivirus Vector for Safe and Efficient In Vivo Gene Delivery,J.Virol.(1998)72(12):9873-9880)。取决于以上提到的基于病毒载体的疫苗平台的包装能力，此方法可以递送编码一个或多个新抗原肽的一个或多个核苷酸序列。这些序列可以侧接非突变序列，可以由连接子分开，或者可以在前面具有一个或多个靶向亚细胞区室的序列(参见例如，Gros等人,Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients,Nat Med.(2016)22(4):433-8;Stronen等人,Targeting of cancer neoantigens with donor-derived T cell receptor repertoires,Science.(2016)352(6291):1337-41;Lu等人，Efficient identification of mutated cancer antigens recognized by T cells

associated with durable tumor regressions,Clin Cancer Res.(2014)20(13):3401-10)。在引入宿主中后,受感染的细胞表达新抗原,并由此引起针对肽的宿主免疫(例如CTL)反应。可用于免疫方案的牛痘载体和方法描述于例如美国专利第4,722,848号中。另一载体是卡介苗(Bacille Calmette Guerin,BCG)。BCG载体描述于Stover等人(Nature 351:456-460(1991))中。根据本文的描述,本领域技术人员将显而易见可用于新抗原的治疗性施用或免疫的多种其他疫苗载体,例如,伤寒沙门氏菌(Salmonella typhi)载体。

[0239] 施用核酸的方式使用了编码一个或多个表位的微型基因构建体。为了产生用于在人细胞中表达的编码所选CTL表位的DNA序列(微型基因),对这些表位的氨基酸序列进行逆翻译。使用人密码子用法表指导各氨基酸的密码子选择。将这些表位编码DNA序列直接邻接,产生连续多肽序列。为了优化表达和/或免疫原性,可以将另外的元件并入微型基因设计中。可以被逆翻译并且包括在微型基因序列中的氨基酸序列的实例包括:辅助T淋巴细胞、表位、前导(信号)序列和内质网滞留信号。此外,通过邻近CTL表位包括合成(例如聚丙氨酸)或天然存在的侧接序列可以改善CTL表位的MHC呈递。通过组装编码微型基因正链和负链的寡核苷酸,将微型基因序列转化成DNA。使用众所周知的技术,在适当条件下合成、磷酸化、纯化重叠寡核苷酸(30-100个碱基长)并使其退火。使用T4 DNA连接酶接合寡核苷酸的末端。接着,可以将这一编码CTL表位多肽的合成微型基因克隆至所希望的表达载体中。

[0240] 可以使用多种配制物制备注射用纯化质粒DNA。这些方法中最简单的方法是在无菌磷酸盐缓冲生理盐水(PBS)中使冻干的DNA复水。多种方法已有描述,并且新技术也可以使用。如上文所述,核酸宜用阳离子性脂质配制。此外,还可以使统称为保护性、相互作用性、非缩合性(PINC)的糖酯、促融脂质体、肽和化合物与纯化的质粒DNA形成复合物以影响各种变量,如稳定性、肌肉内分散或向特定器官或细胞类型的运输。

[0241] 还公开了一种制造肿瘤疫苗的方法,所述方法包括执行本文所公开的方法的各个步骤;以及产生包含多个新抗原或所述多个新抗原的子集的肿瘤疫苗。

[0242] 本文所公开的新抗原可以使用本领域中已知的方法制造。举例来说,本文所公开的产生新抗原或载体(例如包括至少一个编码一个或多个新抗原的序列的载体)的方法可以包括在适于表达所述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个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、电泳技术、免疫技术、沉淀、透析、过滤、浓缩和等电聚焦技术。

[0243] 宿主细胞可以包括中国仓鼠卵巢(CHO)细胞、NSO细胞、酵母或HEK293细胞。宿主细胞可以用一个或多个多核苷酸转化,所述一个或多个多核苷酸包含至少一个编码本文所公开的新抗原或载体的核酸序列,任选地其中分离的多核苷酸另外包含可操作地连接到所述至少一个编码新抗原或载体的核酸序列的启动子序列。在某些实施方案中,所述分离的多核苷酸可以是cDNA。

[0244] VII.新抗原鉴别

[0245] VII.A.新抗原候选物的鉴别。

[0246] 有关以NGS分析肿瘤和正常外显子组和转录组的研究方法已有描述且被应用于新抗原鉴别邻域中。[6,14,15]以下实施例考虑了在临床环境中对于新抗原鉴别具有较高灵敏度和特异性的某些优化措施。这些优化措施可以分为两个领域,即与实验室方法有关的优化和与NGS数据分析有关的优化。

[0247] VII.A.1.实验室方法优化

[0248] 此处提出的方法改进通过将所开发的有关可靠地评估靶癌症组中的癌症驱动基因的概念16扩展至新抗原鉴别所需的全外显子组和全转录组环境，解决了从肿瘤含量较低并且体积较小的临床试样中高准确性发现新抗原的难题。确切地说，这些改进包括：

[0249] 1.靶向整个肿瘤外显子组的深度(>500×)独特平均覆盖率，以检测由于肿瘤含量低或处于亚克隆状态而以低突变等位基因频率存在的突变。

[0250] 2.靶向整个肿瘤外显子组的均匀覆盖率，其中在<100×下覆盖<5％的碱基，由此通过例如以下方式使遗漏新抗原的可能性最低：

[0251] a.采用基于DNA的捕捉探针和个别探针QC[17]

[0252] b.包括针对覆盖较少的区域的额外诱饵

[0253] 3.靶向整个正常外显子组的均匀覆盖率，其中在<20×下覆盖<5％的碱基，由此对于体细胞/生殖系状态可能有最少的新抗原未被分类(并因此不能用作TSNA)

[0254] 4.为了使需要测序的总量减到最少，序列捕捉探针应被设计成仅针对基因编码区，因为非编码RNA不会产生新抗原。其它优化包括：

[0255] a.针对HLA基因的补充探针，这些基因富含GC并且通过标准外显子组测序很难捕捉[18]

[0256] b.排除由于如表达水平不足、蛋白酶体消化欠佳或不常见的序列特征等因素而被预测产生极少或不产生候选新抗原的基因。

[0257] 5.肿瘤RNA将通常同样在高深度(>100M个读段)下测序，以便能够进行变体检测、基因和剪接变体(“同功型”)表达水平的定量，以及融合物检测。来自FFPE样品的RNA将使用基于探针的富集方法19，使用与捕捉DNA中的外显子组相同或类似的探针进行提取。

[0258] VII.A.2.NGS数据分析优化

[0259] 分析方法的改进解决了常用研究突变调用方法灵敏度和特异性欠佳的问题，并且特别考虑到了在临床环境中与新抗原鉴别相关的定制。这些包括：

[0260] 1.使用HG38参考人基因组或后续版本进行比对，因为相对于先前的基因组版本，所述基因组含有多个MHC区域组装体，较佳地反映了群体多态性。

[0261] 2.通过合并由不同程序得到的结果[5]，克服单个变体调用程序的局限性[20]

[0262] a.利用一套工具，检测肿瘤DNA、肿瘤RNA及正常DNA中的单核苷酸变体和插入缺失，所述套工具包括：基于肿瘤与正常DNA的比较的程序，如Strelka[21]和Mutect[22]；和并入了肿瘤DNA、肿瘤RNA及正常DNA的程序，如UNCeqR，特别适用于低纯度样品[23]。

[0263] b.插入缺失将利用执行局部再组装的程序测定，如Strelka和ABRA[24]。

[0264] c.结构重排将使用专用工具测定，如Pindel[25]或Breakseq[26]。

[0265] 3.为了检测并防止样品调换，将在选定的多态性位点数量下，比较来自同一患者的样品中的变体调用。

[0266] 4.针对伪调用的广泛过滤将例如通过以下方式进行：

[0267] a.移除在正常DNA中发现的变体，在低覆盖率下可能使用不严格的检测参数，并且在插入缺失情况下使用容许的接近标准

[0268] b.移除由低定位质量或低碱基质量引起的变体[27]。

[0269] c.移除来源于反复出现的测序伪影的变体，即使在相应的正常情况下未观察到[27]。

实例包括主要在一条链上检测到的变体。

[0270] d.移除不相关的对照物集合中检测到的变体[27]

[0271] 5.使用seq2HLA[28]、ATHLATES[29]或Optitype之一，从正常外显子组中准确地调用HLA，并且还将外显子组与RNA测序数据组合[28]。其它可能的优化包括采用专用于HLA分型的分析，如长读段DNA测序[30]，或调适用于接合RNA片段的方法以保持连续性[31]。

[0272] 6.针对由肿瘤特异性剪接变体产生的neo-ORF的稳健检测将通过使用CLASS[32]、Bayesembler[33]、StringTie[34]或类似程序以其参考引导的模式，根据RNA-seq数据组装转录物来进行(即，使用已知的转录物结构而非尝试在每个实验中重新构建整个转录物)。尽管Cufflinks[35]通常被用于此目的，但它常常会不合情理地产生大量剪接变体，其中有许多比全长基因要短得多，并且无法回收简单的阳性对照。编码序列和无义介导的衰变可能性将通过如SpliceR[36]和MAMBA[37]等工具，利用重新引入的突变序列测定。基因表达将利用如Cufflinks[35]或Express(Roberts和Pachter,2013)等工具测定。野生型和突变体特异性表达计数和/或相对水平将利用开发用于这些目的的工具，如ASE[38]或HTSeq[39]测定。可能的过滤步骤包括：

[0273] a.移除被认为表达不足的候选neo-ORF。

[0274] b.移除被预测会触发无义介导的衰变(NMD)的候选neo-ORF。

[0275] 7.仅在RNA中观察到的无法直接验证为肿瘤特异性抗原的候选新抗原(例如neoORF)将根据额外参数，例如通过考虑以下因素而归类为可能是肿瘤特异性的：

[0276] a.存在仅支持肿瘤DNA的顺式作用移码或剪接位点突变

[0277] b.在剪接因子中存在仅证实肿瘤DNA的反式作用突变。举例来说，在利用R625突变型SF3B1进行的三个独立公布的实验中，尽管一个实验检查到葡萄膜黑素瘤患者[40]，第二个实验检查到葡萄膜黑素瘤细胞系[41]，而第三个实验检查到乳癌患者[42]，但展现最大剪接差异的基因是一致的。

[0278] c.对于新剪接同功型，在RNASeq数据中存在确证的“新”剪接-接合读段。

[0279] d.对于新重排，有确证在肿瘤DNA中存在而在正常DNA中不存在的近似外显子读段

[0280] e.基因表达概略中缺乏，如GTEx[43](即，使得不太可能为生殖系起源)

[0281] 8.通过直接比较组装的DNA肿瘤与正常读段(或来自这些读段的k-mer)来补充基于参考基因组比对的分析以避免基于比对和注释的错误和伪影。(例如对于在生殖系变体或重复序列插入缺失附近出现的体细胞变体)

[0282] 在具有聚腺苷酸化RNA的样品中，RNA-seq数据中病毒和微生物RNA的存在将使用RNA CoMPASS[44]或类似方法评估，以鉴别可以预测患者响应的其它因素。

[0283] VII.B.HLA肽的分离和检测

[0284] HLA-肽分子的分离在溶胞和溶解组织样品之后，使用经典免疫沉淀(IP)方法进行[55-58]。使用澄清的溶解产物进行HLA特异性IP。

[0285] 免疫沉淀是使用偶合至珠粒的抗体进行，其中所述抗体对HLA分子具有特异性。对于全I类HLA免疫沉淀，使用全I类CR抗体，对于II类HLA-DR，使用HLA-DR抗体。在过夜培育期间，将抗体共价连接至NHS-琼脂糖珠粒。在共价连接后，洗涤珠粒并等分试样用于IP。[59,60]免疫沉淀也可以使用未共价结合至磁珠的抗体进行。通常，使用包被有蛋白A和/或蛋白G的琼脂糖或磁珠将抗体固定在色谱柱上来完成此操作。下面列出了一些可用于选择性富集MHC/

肽复合物的抗体。

[0286]

抗体名称	特异性
W6/32	I类HLA-A,B,C
L243	II类-HLA-DR
Tu36	II类-HLA-DR
LN3	II类-HLA-DR
Tu39	II类-HLA-DR,DP,DQ

[0287] 将澄清的组织溶解产物添加至抗体珠粒中进行免疫沉淀。免疫沉淀后,从溶解产物移除珠粒,并储存溶解产物用于另外的实验,包括另外的IP。洗涤IP珠粒以移除非特异性结合并使用标准技术,从珠粒洗脱下HLA/肽复合物。使用分子量旋转柱或C18分级分离,从肽移除蛋白质组分。通过SpeedVac蒸发使所得肽变干并且在一些情形中在-20C下储存以待MS分析。

[0288] 干燥的肽在适于反相色谱法的HPLC缓冲液中复水并装载至C-18微毛细管HPLC柱上以在Fusion Lumos质谱仪(Thermo)中进行梯度洗脱。在Orbitrap检测器中在高分辨率下收集肽质/荷比(m/z)的MS1谱,随后在所选离子经历HCD片段化后,在离子阱检测器中收集MS2低分辨率扫描谱。另外,可以使用CID或ETD片段化方法,或三种技术的任何组合获得MS2谱,以达到所述肽的较高氨基酸覆盖率。还可以在Orbitrap检测器中用高分辨率质量精度测量MS2谱。

[0289] 使用Comet[61,62],针对蛋白质数据库搜索由各分析得到的MS2谱并使用Percolator[63-65]对肽鉴别进行评分。可以使用PEAKS studio(Bioinformatics Solutions Inc.)进行另外的测序,并且可以使用其他搜索引擎或其他测序方法,包括光谱匹配和从头测序[75]。

[0290] VII.B.1.支持全面HLA肽测序的MS检测限研究。

[0291] 使用肽YVYVADVAAK,利用装载至LC柱上的不同量的肽确定检测限。测试肽的量是1pmol、100fmol、10fmol、1fmol和100amol。(表1)结果显示于图1F中。这些结果表明,最低检测限(LoD)是埃摩尔(attomol)范围(10-18),动态范围跨五个数量级,并且信噪比看来足以在低飞摩尔(femtomol)范围(10-15)内进行测序。

[0292]

肽 m/z	装载于柱上	在 1e9 个细胞中的拷贝数/细胞
566.830	1 pmol	600
562.823	100 fmol	60
559.816	10 fmol	6
556.810	1 fmol	0.6
553.802	100 amol	0.06

[0293]

[0294] VIII.呈递模型

[0295] VIII.A.系统概述

[0296] 图2A是根据一个实施方案,用于鉴别患者体内肽呈递的可能性的环境100的概述。环境100提供背景以便引入呈递鉴别系统160,所述系统本身包括呈递信息存储器165。

[0297] 呈递鉴别系统160是一个或多个在如以下关于图14所论述的计算系统中体现的计

算机模型，其接收与MHC等位基因集合有关的肽序列并测定这些肽序列将被所述相关MHC等位基因集合中的一个或多个MHC等位基因呈递的可能性。呈递鉴别系统160可以应用于I类和II类MHC等位基因两者。这在多种情形中都适用。呈递鉴别系统160的一个具体使用情形是，它能够接收与来自患者110的肿瘤细胞的MHC等位基因集合有关的候选新抗原的核苷酸序列，并测定这些候选新抗原将被所述肿瘤的相关MHC等位基因中的一个或多个呈递和/或在患者110的免疫系统中诱导免疫原性反应的可能性。可以选出被系统160测定具有高可能性的候选新抗原用于包括在疫苗118中，此类抗肿瘤免疫反应可以由提供肿瘤细胞的患者110的免疫系统引发。

[0298] 呈递鉴别系统160通过一个或多个呈递模型测定呈递可能性。确切地说，呈递模型生成给定肽序列是否将由相关MHC等位基因集合呈递的可能性，并且这是基于存储在存储器165中的呈递信息生成的。举例来说，呈递模型可以生成肽序列“YVYVADVAAK”是否将由等位基因HLA-A*02:01、HLA-A*03:01、HLA-B*07:02、HLA-B*08:03、HLA-C*01:04的集合呈递于样品的细胞表面上的可能性。呈递信息165含有关于肽是否结合至不同类型的MHC等位基因以使得这些肽被MHC等位基因呈递的信息，所述信息在模型中是根据肽序列中氨基酸的位置确定。呈递模型可以基于呈递信息165预测未被识别的肽序列的呈递是否会与相关MHC等位基因集合相关联。如前所述，呈递模型可以应用于I类和II类MHC等位基因两者。

[0299] VIII.B.呈递信息

[0300] 图2示出了根据一个实施方案的获得呈递信息的方法。呈递信息165包括两个通用信息类别：等位基因相互作用信息和等位基因非相互作用信息。等位基因相互作用信息包括影响与MHC等位基因的类型相关的肽序列的呈递的信息。等位基因非相互作用信息包括影响与MHC等位基因的类型无关的肽序列的呈递的信息。

[0301] VIII.B.1.等位基因相互作用信息

[0302] 等位基因相互作用信息主要包括经过鉴别的肽序列，已知这些肽序列已经被来自人、小鼠等的一个或多个经过鉴别的MHC分子呈递。值得注意的是，这可能包括或可能不包括从肿瘤样品获得的数据。可以从表达单个MHC等位基因的细胞鉴别出所呈递的肽序列。在这一情形中，所呈递的肽序列一般是从单个等位基因细胞系收集，这些细胞系被工程改造成表达预定MHC等位基因并且随后暴露于合成蛋白质。在MHC等位基因上呈递的肽是通过如酸洗脱等技术分离并通过质谱法鉴别。图2B示出了这一情形的一个实施例，其中分离出在预定MHC等位基因HLA-DRB1*12:01上呈递的示例性肽YEMFNDKSQRAPDDKMF并通过质谱法鉴别。由于在此情况下，肽是通过被工程改造成表达单一预定MHC蛋白质的细胞鉴别，故呈递的肽与其所结合的MHC蛋白质之间的直接关联是确定已知的。

[0303] 也可以从表达多个MHC等位基因的细胞收集所呈递的肽序列。通常，在人体中，一种细胞表达6种不同类型的MHC-I和至多12种不同类型的MHC-II分子。如此呈递的肽序列可以从被工程改造成表达多个预定MHC等位基因的多等位基因细胞系鉴别到。还可以从组织样品，如正常组织样品或肿瘤组织样品鉴别如此呈递的肽序列。特别就这一情形来说，MHC分子可以从正常或肿瘤组织免疫沉淀。在多个MHC等位基因上呈递的肽可类似地通过如酸洗脱等技术分离并通过质谱法鉴别。图2C示出了此种情形的一个实施例，其中将六个示例性肽YEMFNDKSF、HROEIFSHDFJ、FJIEJFOESS、NEIOREIREI、JFKSIFEMMSJDSSUIFLKSJFIEIFJ和KNFLENFIESOFI呈递于所鉴别的I类MHC等位基因HLA-A*01:01、HLA-A*02:01、HLA-B*07:02、

HLA-B*08:01和II类MHC等位基因HLA-DRB1*10:01、HLA-DRB1:11:01并且分离，并通过质谱法鉴别。相对于单等位基因细胞系，呈递的肽与其所结合的MHC蛋白质之间的直接关联可能是未知的，因为结合肽是在鉴别之前与MHC分子分离。

[0304] 等位基因相互作用信息还可以包括质谱离子流，其取决于肽-MHC分子复合物的浓度和肽电离效率。电离效率以序列依赖性方式随肽而变化。一般来说，电离效率随肽而在约两个数量级内变化，而肽-MHC复合物的浓度在比其更大的范围内变化。

[0305] 等位基因相互作用信息还可以包括给定MHC等位基因与给定肽之间结合亲和力的测量或预测。(72,73,74)一个或多个亲和力模型可以生成此类预测。举例来说，再看回图1D中所示的实施例，呈递信息165可以包括肽YEMFNDKSF与等位基因I类HLA-A*01:01之间的1000nM的结合亲和力预测值。IC50>1000nm的肽很少被MHC呈递，且较低的IC50值使呈递机率增加。呈递信息165可以包括肽KNFLENFIESOFI和II类等位基因HLA-DRB1:11:01之间的结合亲和力预测。

[0306] 等位基因相互作用信息也可以包括所述MHC复合物稳定性的测量或预测。一个或多个稳定性模型可以生成此类预测。较稳定的肽-MHC复合物(即，半衰期较长的复合物)比较可能在肿瘤细胞上和在遭遇疫苗抗原的抗原呈递细胞上以高拷贝数呈递。举例来说，再看回图2C中所示的实施例，呈递信息165可以包括I类分子HLA-A*01:01的半衰期是1小时的稳定性预测值。呈递信息165可以包括II类分子HLA-DRB1:11:01的半衰期的稳定性预测值。

[0307] 等位基因相互作用信息也可以包括测量或预测的肽-MHC复合物的形成反应速率。以较高速率形成的复合物比较可能以高浓度呈递于细胞表面上。

[0308] 等位基因相互作用信息还可以包括肽的序列和长度。I类MHC分子通常偏好呈递长度介于8与15个肽之间的肽。所呈递的肽中有60-80％的长度是9个。II类MHC分子通常更优先呈递介于6到30个肽之间的肽。

[0309] 等位基因相互作用信息还可以包括新抗原编码肽上激酶序列基元的存在，以及新抗原编码肽上特定翻译后修饰的不存在或存在。激酶基元的存在会影响翻译后修饰的机率，所述翻译后修饰可能增强或干扰MHC结合。

[0310] 等位基因相互作用信息还可以包括翻译后修饰过程中所涉及的蛋白质，例如激酶的表达水平或活性水平(如由RNA seq、质谱法或其他方法所测量或预测)。

[0311] 等位基因相互作用信息还可以包括来自表达特定MHC等位基因的其他个体的细胞中具有相似序列的肽的呈递机率，这可通过质谱蛋白组学或其他手段评估。

[0312] 等位基因相互作用信息还可以包括所讨论的个体中特定MHC等位基因的表达水平(例如，如通过RNA-seq或质谱法测量)。相较于最强地结合至以低水平表达的MHC等位基因的肽，最强地结合至以高水平表达的MHC等位基因的肽比较可能被呈递。

[0313] 等位基因相互作用信息还可以包括不依赖于总体新抗原编码肽序列而在表达特定MHC等位基因的其他个体中由特定MHC等位基因呈递的机率。

[0314] 等位基因相互作用信息还可以包括不依赖于总体肽序列而在其他个体中由同一家族分子(例如HLA-A、HLA-B、HLA-C、HLA-DQ、HLA-DR、HLA-DP)中的MHC等位基因呈递的机率。举例来说，HLA-C分子的表达水平通常低于HLA-A或HLA-B分子，且由此可推断，由HLA-C呈递肽的机率低于由HLA-A或HLA-B呈递的机率。再举一个例子，HLA-DP的表达水平通常低于HLA-DR或HLA-DQ，且由此可推断，由HLA-DP呈递肽的机率低于由HLA-DR或HLA-DQ呈递的

机率。

[0315] 等位基因相互作用信息还可以包括特定MHC等位基因的蛋白质序列。

[0316] 以下部分中所列的任何MHC等位基因非相互作用信息也可以按MHC等位基因相互作用信息的方式进行建模。

[0317] VIII.B.2.等位基因非相互作用信息

[0318] 等位基因非相互作用信息可以包括在源蛋白质序列内侧接新抗原编码肽的C末端序列。对于MHC-I,C末端侧接序列可能影响肽的蛋白酶体加工。不过,C末端侧接序列是在肽转运至内质网并遇到细胞表面上的MHC等位基因之前,在蛋白酶体作用下自所述肽裂解得到。因此,MHC分子接收不到有关C末端侧接序列的信息,且由此,C末端侧接序列的影响不会随MHC等位基因类型而变化。举例来说,再参看图FIG.2C中所示的实施例,呈递信息165可以包括从肽的源蛋白鉴别到的呈递肽FJIEJFOESS的C末端侧接序列FOEIFNDKSLDKFJI。

[0319] 等位基因非相互作用信息也可以包括mRNA定量测量。举例来说,可以获得与提供质谱训练数据相同的样品的mRNA定量数据。如稍后参照图13H所描述,RNA表达水平被鉴别为肽呈递的强预测因子。在一个实施方案中,mRNA定量测量值是由软件工具RSEM鉴别得到。有关RSEM软件工具的详细实施方式可见于Bo Li和Colin N.Dewey.RSEM:accurate transcript quantification from RNA-Seq data with or without a reference genome.BMC Bioinformatics,12:323,2011年8月。在一个实施方案中,mRNA定量是以每一百万条定位读段数中每千碱基转录物的片段数(FPKM)为单位度量。

[0320] 等位基因非相互作用信息还可以包括在源蛋白质序列内侧接所述肽的N末端序列。

[0321] 等位基因非相互作用信息还可以包括肽序列的源基因。可以将源基因定义为肽序列的Ensembl蛋白家族。在另一些例子中,源基因可以被定义为肽序列的源DNA或源RNA。可以例如将源基因表示为编码蛋白质的一串核苷酸,或者基于已知编码特定蛋白质的已知DNA或RNA序列的命名集合将更直接地表示。在另一个例子中,等位基因非相互作用信息还可以包括从数据库如Ensembl或RefSeq中提取的肽序列的源转录本或同工型或潜在的源转录本或同工型的集合。

[0322] 等位基因非相互作用信息还可以包括肽序列来源的细胞的组织类型、细胞类型或肿瘤类型。

[0323] 等位基因非相互作用信息还可以包括在所述肽中蛋白酶裂解基元的存在,任选地根据肿瘤细胞中相应蛋白酶的表达(如通过RNA-seq或质谱法测量)加权。含有蛋白酶裂解基元的肽不太可能被呈递,因为这些肽比较容易被蛋白酶降解,并因此在细胞内不太稳定。

[0324] 等位基因非相互作用信息还可以包括如在适当细胞类型中测量的源蛋白的转换率。转换率较快(即,半衰期较短)会增加呈递机率;不过,如果在不相似的细胞类型中测量,则此特征的预测能力较低。

[0325] 等位基因非相互作用信息还可以包括如通过RNA-seq或蛋白质组质谱法所测量,或如根据在DNA或RNA序列数据中检测到的生殖系或体细胞剪接突变的注释所预测的源蛋白的长度,任选地考虑在肿瘤细胞中表达水平最高的特定剪接变体(“同功型”)。

[0326] 等位基因非相互作用信息还可以包括肿瘤细胞中蛋白酶体、免疫蛋白酶体、胸腺蛋白酶体或其他蛋白酶的表达水平(可以通过RNA-seq、蛋白质组质谱法或免疫组织化学分

析测量)。不同的蛋白酶体具有不同的裂解位点偏好。与表达水平成比例的各类型蛋白酶体的裂解偏好将被给予较大权重。

[0327] 等位基因非相互作用信息还可以包括肽的源基因的表达水平(例如通过RNA-seq或质谱法测量)。可能的优化措施包括调整表达水平测量值以说明肿瘤样品内基质细胞和肿瘤浸润淋巴细胞的存在。来自表达水平较高的基因的肽比较可能被呈递。来自表达水平不可检测的基因的肽可以不予考虑。

[0328] 等位基因非相互作用信息还可以包括如由无义介导的衰变模型,例如来自Rivas等人,Science 2015的模型所预测的新抗原编码肽的源mRNA将经历无义介导的衰变的机率。

[0329] 等位基因非相互作用信息还可以包括在各种细胞周期阶段期间肽的源基因的典型肿瘤特异性表达水平。以总体较低水平表达(如通过RNA-seq或质朴蛋白质组学所测量)但已知在特定细胞周期阶段期间高水平表达的基因所产生的呈递肽可能多于以极低水平稳定表达的基因。

[0330] 等位基因非相互作用信息还可以包括例如uniProt或PDB http://www.rcsb.org/pdb/home/home.do中提供的源蛋白特征的综合目录。这些特征尤其可以包括:蛋白质的二级和三级结构、亚细胞定位11、基因本体(Gene ontology,GO)项。确切地说,这一信息可以含有在蛋白质水平上起作用的注释,例如5'UTR长度;以及在特定残基水平上起作用的注释,例如在残基300与310之间的螺旋基元。这些特征还可以包括转角基元、折叠基元和无序残基。

[0331] 等位基因非相互作用信息还可以包括描述含有所述肽的源蛋白的结构域的特性的特征,例如:二级或三级结构(例如α螺旋对比β折叠);选择性剪接。

[0332] 等位基因非相互作用信息还可以包括描述在所述肽的源蛋白中所述肽的位置处存在或不存在呈递热点的特征。

[0333] 等位基因非相互作用信息还可以包括其他个体中来自相关肽的源蛋白的肽的呈递机率(在调整这些个体中源蛋白的表达水平和这些个体的不同HLA类型的影响之后)。

[0334] 等位基因非相互作用信息还可以包括由于技术偏差而无法通过质谱法检测到或过量表示所述肽的机率。

[0335] 通过基因表达测定如RNASeq、微阵列、靶向组如Nanostring所测量的各种基因模块/路径的表达,或通过如RT-PCR等测定(无需含有所述肽的源蛋白)所测量的基因模块的单基因/多基因代表提供了有关肿瘤细胞、基质或肿瘤浸润淋巴细胞(TIL)的状态的信息。

[0336] 等位基因非相互作用信息还可以包括肿瘤细胞中肽的源基因的拷贝数。举例来说,在肿瘤细胞中经历纯合子缺失的基因的肽可以指定为呈递机率是零。

[0337] 等位基因非相互作用信息还可以包括肽结合至TAP的机率或肽与TAP的结合亲和力测量值或预测值。比较可能结合至TAP的肽,或以较高亲和力结合TAP的肽比较可能被MHC-I呈递。

[0338] 等位基因非相互作用信息还可以包括肿瘤细胞中TAP的表达水平(可以通过RNA-seq、蛋白质组质谱法、免疫组织化学分析测量)。对于MHC-I,较高的TAP表达水平会增加所有肽的呈递机率。

[0339] 等位基因非相互作用信息还可以包括肿瘤突变的存在或不存在,包括但不限于:

[0340] i.已知癌症驱动基因,如EGFR、KRAS、ALK、RET、ROS1、TP53、CDKN2A、CDKN2B、NTRK1、NTRK2、NTRK3中的驱动突变

[0341] ii.编码抗原呈递机器中所涉及的蛋白质的基因(例如B2M、HLA-A、HLA-B、HLA-C、TAP-1、TAP-2、TAPBP、CALR、CNX、ERP57、HLA-DM、HLA-DMA、HLA-DMB、HLA-DO、HLA-DOA、HLA-DOBHLA-DP、HLA-DPA1、HLA-DPB1、HLA-DQ、HLA-DQA1、HLA-DQA2、HLA-DQB1、HLA-DQB2、HLA-DR、HLA-DRA、HLA-DRB1、HLA-DRB3、HLA-DRB4、HLA-DRB5或编码蛋白酶体或免疫蛋白酶体的组分的任何基因)中的突变。呈递依赖于肿瘤中经历功能丧失性突变的抗原呈递机器组分的肽具有降低的呈递机率。

[0342] 存在或不存在功能性生殖系多态现象,包括但不限于:

[0343] i.编码抗原呈递机器中所涉及的蛋白质的基因(例如B2M、HLA-A、HLA-B、HLA-C、TAP-1、TAP-2、TAPBP、CALR、CNX、ERP57、HLA-DM、HLA-DMA、HLA-DMB、HLA-DO、HLA-DOA、HLA-DOBHLA-DP、HLA-DPA1、HLA-DPB1、HLA-DQ、HLA-DQA1、HLA-DQA2、HLA-DQB1、HLA-DQB2、HLA-DR、HLA-DRA、HLA-DRB1、HLA-DRB3、HLA-DRB4、HLA-DRB5或编码蛋白酶体或免疫蛋白酶体的组分的任何基因)中的功能性生殖系多态现象

[0344] 等位基因非相互作用信息还可以包括肿瘤类型(例如NSCLC、黑素瘤)。

[0345] 等位基因非相互作用信息还可以包括HLA等位基因的已知功能,如由例如HLA等位基因的后缀所反映。举例来说,等位基因名称HLA-A*24:09N中的N后缀指示未表达并因此不可能呈递表位的无效等位基因;完整HLA等位基因后缀命名法描述于https://www.ebi.ac.uk/ipd/imgt/hla/nomenclature/suffixes.html。

[0346] 等位基因非相互作用信息还可以包括临床肿瘤亚型(例如鳞状肺癌对比非鳞状肺癌)。

[0347] 等位基因非相互作用信息也可以包括吸烟史。

[0348] 等位基因非相互作用信息还可以包括晒伤史、太阳曝晒史或暴露于其他诱变剂的历史。

[0349] 等位基因非相互作用信息还可以包括肽的源基因在相关肿瘤类型或临床亚型中的典型表达,任选地利用驱动基因突变分层。通常在相关肿瘤类型中高水平表达的基因比较可能被呈递。

[0350] 等位基因非相互作用信息还可以包括所有肿瘤中,或同一类型肿瘤中,或来自具有至少一个共有MHC等位基因的个体的肿瘤中,或具有至少一个共有MHC等位基因的个体体内的同一类型肿瘤中的突变频率。

[0351] 就突变的肿瘤特异性肽而言,用于预测呈递机率的特征清单也可以包括突变注释(例如错义、通读、移码突变、融合等)或预测所述突变是否会引起无义介导的衰变(NMD)。举例来说,来自因纯合子早期终止突变而在肿瘤细胞中不翻译的蛋白质区段的肽可以指定为呈递机率是零。NMD使mRNA翻译减少,由此降低呈递机率。

[0352] VIII.C.呈递鉴别系统

[0353] 图3是示出根据一个实施方案的呈递鉴别系统160的计算机逻辑组件的高级框图。在该示例性实施方案中,呈递鉴别系统160包括数据管理模块312、编码模块314、训练模块316和预测模块320。呈递鉴别系统160还包括训练数据存储器170和呈递模型存储器175。所述模型管理系统160的一些实施方案具有与此处所描述不同的模块。类似地,这些模块的功

能分布可能不同于此处描述的模块。

[0354] VIII.C.1.数据管理模块

[0355] 数据管理模块312根据呈递信息165生成数组训练数据170。每组训练数据含有多个数据实例，其中每个数据实例i含有自变量集合z^i，这些自变量包括至少一个呈递或不呈递肽序列p^i、一个或多个与所述肽序列p^i相关联的相关MHC等位基因a^i；和一个因变量y^i，所述因变量表示呈递鉴别系统160有意预测自变量的新值的信息。

[0356] 在本说明书其余部分通篇提到的一个特定的实施方式中，因变量y^i是一种二元标记，指示肽p^i是否被所述一个或多个相关MHC等位基因a^i呈递。不过，应理解，在其他实施方式中，取决于自变量z^i，因变量y^i可以表示呈递鉴别系统160有意进行预测的任何其他类别的信息。举例来说，在另一个实施方式中，因变量y^i还可以是指示所鉴别的数据实例的质谱离子电流的数值。

[0357] 数据实例i的肽序列p^i是具有k_i个氨基酸的序列，其中k_i可以在随数据实例i而在一定范围内变化。举例来说，该范围对于I类MHC可以是8-15，或对于II类MHC是6-30。在系统160的一个具体实施方式中，一个训练数据集中的所有肽序列p^i可以具有相同长度，例如9。肽序列中氨基酸的数量可以取决于MHC等位基因的类型(例如人体中的MHC等位基因等)而变化。数据实例i的MHC等位基因a^i指示存在的与相应肽序列p^i相关的MHC等位基因。

[0358] 数据管理模块312还可以包括另外的等位基因相互作用变量，如与训练数据170中所包含的肽序列p^i和相关MHC等位基因a^i有关的结合亲和力b^i和稳定性预测值s^i。举例来说，训练数据170可以含有肽p^i与以a^i指示的各相关MHC分子之间的结合亲和力预测值b^i。又如，训练数据170可以含有以a^i指示的各MHC等位基因的稳定性预测值s^i。

[0359] 数据管理模块312还可以包括等位基因非相互作用变量w^i，如与肽序列p^i有关的C末端侧接序列和mRNA定量测量值。

[0360] 数据管理模块312还鉴别不被MHC等位基因呈递的肽序列，以生成训练数据170。一般来说，这涉及在呈递之前，鉴别包括呈递肽序列在内的源蛋白的“较长”序列。当呈递信息含有工程改造的细胞系时，数据管理模块312鉴别这些细胞所暴露的合成蛋白质中未呈递于细胞的MHC等位基因上的一系列肽序列。当呈递信息含有组织样品时，数据管理模块312鉴别作为呈递肽序列的来源的源蛋白，并且鉴别源蛋白中未呈递于组织样品细胞的MHC等位基因上的一系列肽序列。

[0361] 数据管理模块312还可以利用随机氨基酸序列人工产生肽，并将所产生的序列鉴别为不呈递于MHC等位基因上的肽。这可以通过随机产生肽序列实现，使得数据管理模块312能够容易地生成大量有关不呈递于MHC等位基因上的肽的合成数据。由于实际上，只有少量肽序列被MHC等位基因呈递，故合成产生的肽序列很有可能不会被MHC等位基因呈递，即使这些序列被包括在细胞加工的蛋白质中。

[0362] 图4示出根据一个实施方案的示例性训练数据集170A。确切地说，训练数据170A中的前3个数据实例指示由包含等位基因HLA-C*01:03以及3个肽序列QCEIOWAREFLKEIGJ、FIEUHFWI和FEWRHRJTRUJR的单等位基因细胞系得到的肽呈递信息。训练数据170A中的第四个数据实例指示由包含等位基因HLA-B*07:02、HLA-C*01:03、HLA-A*01:01和一个肽序列QIEJOEIJE的多等位基因细胞系得到的肽信息。第一个数据实例指示，肽序列QCEIOWARE不被等位基因HLA-DRB3:01:01呈递。如前两段所论述，阴性标记的肽序列可以由数据管理模

块312随机产生，或从呈递肽的源蛋白鉴别得到。训练数据170A还包括肽序列-等位基因对的1000nM的结合亲和力预测值以及半衰期是1小时的稳定性预测值。训练数据170A还包括等位基因非相互作用变量，如肽FJELFISBOSJFIE的C末端侧接序列，以及10^2TPM的mRNA定量测量值。第四个数据实例指示，肽序列QIEJOEIJE被等位基因HLA-B*07:02、HLA-C*01:03或HLA-A*01:01之一呈递。训练数据170A还包括各等位基因的结合亲和力预测值和稳定性预测值，以及所述肽的C末端侧接序列和所述肽的mRNA定量测量值。

[0363] VIII.C.2.编码模型

[0364] 编码模块314将训练数据170中所包含的信息编码成可以用于产生一个或多个呈递模型的数字表示。在一个实施方式中，编码模块314在预定的20字母氨基酸字母表内独热编码序列(例如肽序列或C末端侧接序列)。确切地说，具有k_i个氨基酸的肽序列p^i表示为具有$20\cdot k_i$个元素的行向量，其中$p^i_{20\cdot(j-1)+1}, p^i_{20\cdot(j-1)+2}, \cdots, p^i_{20\cdot j}$当中对应于字母表中在肽序列第j位的氨基酸的单一元素的值是1。另外，其余元素的值是0。例如，对于给定的字母表{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}，数据实例i的具有3个氨基酸的肽序列EAF可以由具有60个元素的行向量表示p^i=[0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]。C末端侧接序列c^i，以及MHC等位基因的蛋白质序列d_h和呈递信息中的其他序列数据都可以按与上文所描述类似的方式编码。

[0365] 当训练数据170含有氨基酸长度不同的序列时，编码模块314也可以通过添加PAD字符以扩充预定字母表，将肽编码成相等长度的向量。举例来说，这可以通过用PAD字符对肽序列进行左侧填充，直到所述肽序列的长度达到训练数据170中具有最大长度的肽序列来进行。因此，当具有最大长度的肽序列具有k_{max}个氨基酸时，编码模块314将各序列以数字方式表示为具有$(20+1)\cdot k_{max}$个元素的行向量。例如，对于扩充的字母表{PAD,A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}和$k_{max}=5$的最大氨基酸长度，所述具有3个氨基酸的相同示例性肽序列可以由具有105个元素的行向量表示：p^i=[1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]。C末端侧接序列c^i或其他序列数据可以按与上文描述类似的方式编码。因此，肽序列p^i或c^i的每个自变量或每一列表示在所述序列特定位置处特定氨基酸的存在。

[0366] 尽管以上编码序列数据的方法是参照具有氨基酸序列的序列描述，但所述方法可以类似地扩展至其他类型的序列数据，如DNA或RNA序列数据等。

[0367] 编码模块314还将数据实例i的一个或多个MHC等位基因a^i编码成具有m个元素的行向量，其中每个元素h=1,2,…,m对应于唯一鉴别的MHC等位基因。对应于所鉴别的数据实例i的MHC等位基因的元素的值是1。另外，其余元素的值是0。例如，m=4种唯一鉴别的MHC等位基因类型{HLA-A*01:01,HLA-C*01:08,HLA-B*07:02,HLA-DRB1*10:01}当中对应于多等位基因细胞系的数据实例i的等位基因HLA-B*07:02和HLA-DRB1*10:01可以由具有4个元素的行向量表示：a^i=[0 0 1 1]，其中$a_3^i=1$并且$a_4^i=1$。尽管本文中用4种鉴别的MHC等位基因类型描述所述实施例，但MHC等位基因类型的数量实际上可以是数百或数千。如先前所论述，每个数据实例i通常含有最多6种不同的与肽序列p_i相关的MHC等位基因类型。

[0368] 编码模块314还将每个数据实例i的标记y_i编码为具有来自集合{0,1}的值的二元变量，其中值1指示肽x^i由相关的MHC等位基因a^i中的一个呈递，而值0指示肽x^i不被任何相关的MHC等位基因a^i呈递。当因变量y_i表示质谱离子电流时，编码模块314可以另外使用各种函数，如对[0,∞)之间的离子电流具有(-∞,∞)范围的对数函数等缩放这些值。

[0369] 编码模块314可以将有关肽p_i和相关MHC等位基因h的一对等位基因相互作用变量x_h^i表示为行向量，其中等位基因相互作用变量的数字表示相继地串接。举例来说，编码模块314可以将x_h^i表示为等于$[p^i]$、$[p^i\ b_h^i]$、$[p^i\ s_h^i]$或$[p^i\ b_h^i\ s_h^i]$的行向量，其中b_h^i是肽p_i和相关MHC等位基因h的结合亲和力预测值，并且类似地s_h^i是关于稳定性。或者，等位基因相互作用变量的一个或多个组合可以个别地存储(例如以个别向量或矩阵形式)。

[0370] 在一个实例中，编码模块314通过将结合亲和力的测量值或预测值并入等位基因相互作用变量x_h^i中来表示结合亲和力信息。

[0371] 在一个实例中，编码模块314通过将结合稳定性的测量值或预测值并入等位基因相互作用变量x_h^i中来表示结合稳定性信息。

[0372] 在一个实例中，编码模块314通过将结合缔合速率的测量值或预测值并入等位基因相互作用变量x_h^i中来表示结合缔合速率信息。

[0373] 在一个实例中，对于由I类MHC分子呈递的肽，编码模块314将肽长度表示为向量T_k＝Tk＝[(Lk＝8)(Lk＝9)(Lk＝10)(Lk＝11)(Lk＝12)(Lk＝13)(Lk＝14)(Lk＝15)]，其中是指示函数，并且L_k表示肽p_k的长度。向量T_k可以被包括在等位基因相互作用变量x_h^i中。在另一个实例中，对于由II类MHC分子呈递的肽，编码模块314将肽长度表示为向量T_k＝[(L_k＝6)(L_k＝7)(L_k＝8)(L_k＝9)(L_k＝10)(L_k＝11)(L_k＝12)(L_k＝13)(L_k＝14)(L_k＝15)(L_k＝16)(L_k＝17)(L_k＝18)(L_k＝19)(L_k＝20)(L_k＝21)(L_k＝22)(L_k＝23)(L_k＝24)(L_k＝25)(L_k＝26)(L_k＝27)(L_k＝28)(L_k＝29)(L_k＝30)]，其中是指示函数，并且L_k表示肽p_k的长度。向量T_k可以被包括在等位基因相互作用变量x_h^i中。

[0374] 在一个实例中，编码模块314通过将基于RNA-seq的MHC等位基因表达水平并入等位基因相互作用变量x_h^i中来表示MHC等位基因的RNA表达信息。

[0375] 类似地，编码模块314可以将等位基因非相互作用变量w^i表示为行向量，其中等位基因非相互作用变量的数字表示相继地串接。举例来说，w^i可以是等于$[c^i]$或$[c^i\ m^i\ w^i]$的行向量，其中w^i是除肽p^i的C末端侧接序列和与所述肽相关的mRNA定量测量值m^i外，还表示任何其他等位基因非相互作用变量的行向量。或者，等位基因非相互作用变量的一个或多个组合可以个别地存储(例如以个别向量或矩阵形式)。

[0376] 在一实例中，编码模块314通过将转换率或半衰期并入等位基因非相互作用变量w^i中来表示肽序列的源蛋白的转换率。

[0377] 在一个实例中，编码模块314通过将蛋白质长度并入等位基因非相互作用变量w^i中来表示源蛋白或同功型的长度。

[0378] 在一个实例中，编码模块314通过将包括$\beta1_i$、$\beta2_i$、$\beta5_i$亚单元在内的免疫蛋白酶体特异性蛋白酶体亚单元的平均表达水平并入等位基因非相互作用变量w^i中来表示免疫蛋白酶体的活化情况。

[0379] 在一个实例中，编码模块314通过将源蛋白的丰度并入等位基因非相互作用变量w^i中来表示肽的源蛋白或者肽的基因或转录物的RNA-seq丰度(通过如RSEM等技术以FPKM、

TPM为单位定量)。

[0380] 在一个实例中,编码模块314通过将利用Rivas等人,Science,2015中的模型估计的肽的源转录物会经历无义介导的衰变(NMD)的机率并入等位基因非相互作用变量w^i中来表示此机率。

[0381] 在一个实例中,编码模块314例如通过使用例如路径中每个基因的RSEM,以TPM为单位定量所述路径中基因的表达水平,接着计算所述路径中所有基因的概括统计量,例如平均值,以此表示经RNA-seq评估的基因模块或路径的活化状态。所述平均值可以并入等位基因非相互作用变量w^i中。

[0382] 在一个实例中,编码模块314通过将拷贝数并入等位基因非相互作用变量w^i中来表示源基因的拷贝数。

[0383] 在一个实例中,编码模块314通过将测量的或预测的TAP结合亲和力值(例如以纳摩尔浓度为单位)包括在等位基因非相互作用变量w^i中来表示TAP结合亲和力。

[0384] 在一个实例中,编码模块314通过将利用RNA-seq测量(并利用例如RSEM,以TPM为单位定量)的TAP表达水平包括在等位基因非相互作用变量w^i中来表示TAP表达水平。

[0385] 在一个实例中,编码模块314在等位基因非相互作用变量w^i中将肿瘤突变表示为指示变量的向量(即,如果肽p^k来自具有KRAS G12D突变的样品,则$d^k=1$,否则是0)。

[0386] 在一个实例中,编码模块314将抗原呈递基因中的生殖系多态性表示为指示变量的向量(即,如果肽p^k来自在TAP中具有物种生殖系多态性的样品,则$d^k=1$)。这些指示变量都可以被包括在等位基因非相互作用变量w^i中。

[0387] 在一个实例中,编码模块314根据肿瘤类型(例如NSCLC、黑素瘤、结肠直肠癌等)的字母表将肿瘤类型表示为长度一独热编码的向量。这些独热编码的变量都可以被包括在等位基因非相互作用变量w^i中。

[0388] 在一个实例中,编码模块314通过用不同后缀处理有4个数字的HLA等位基因来表示MHC等位基因后缀。举例来说,出于所述模型的目的,HLA-A*24:09N被认为是与HLA-A*24:09不同的等位基因。或者,由于以N后缀结尾的HLA等位基因不表达,故可以将以N为后缀的MHC等位基因对所有肽的呈递机率设置成零。

[0389] 在一个实例中,编码模块314根据肿瘤亚型(例如肺腺癌、肺鳞状细胞癌等)的字母表将肿瘤亚型表示为长度一独热编码的向量。这些独热编码的变量都可以被包括在等位基因非相互作用变量w^i中。

[0390] 在一个实例中,编码模块314将吸烟史表示为二元指示变量(如果患者有吸烟史,则$d^k=1$,否则是0),所述变量可以包括在等位基因非相互作用变量w^i中。或者,可以根据吸烟严重程度的字母表,将吸烟史编码为长度一独热编码的变量。举例来说,吸烟状态可以在1-5级量表上评级,其中1指示非吸烟者,并且5指示当前多量吸烟者。由于吸烟史主要与肺部肿瘤相关,故当训练有关多种肿瘤类型的模型时,此变量也可以在患者有吸烟史时定义为等于1并且肿瘤类型是肺部肿瘤,否则是零。

[0391] 在一个实例中,编码模块314将晒伤史表示为二元指示变量(如果患者有重度晒伤史,则$d^k=1$,否则是0),所述变量可以包括在等位基因非相互作用变量w^i中。由于重度晒伤主要与黑素瘤相关,故当训练有关多种肿瘤类型的模型时,此变量也可以在患者有重度晒伤史时定义为等于1并且肿瘤类型是黑素瘤,否则是零。

[0392] 在一个实例中，编码模块314通过使用参考数据库如TCGA将有关人基因组中各基因或转录物的特定基因或转录物的表达水平分布表示为表达水平分布的概括统计量(例如平均值、中值)。确切地说，对于肿瘤类型是黑素瘤的样品中的肽p^k，不仅可以将肽p^k的源基因或转录物的基因或转录物表达水平测量值包括在等位基因非相互作用变量w^i中，而且还包括通过TCGA测量的黑素瘤中肽p^k的源基因或转录物的平均和/或中值基因或转录物表达水平。

[0393] 在一个实例中，编码模块314根据突变类型(例如错义突变、移码突变、NMD诱导的突变等)的字母表将突变类型表示为长度一独热编码的变量。这些独热编码的变量都可以被包括在等位基因非相互作用变量w^i中。

[0394] 在一个实例中，编码模块314在等位基因非相互作用变量w^i中将蛋白质的蛋白质水平特征表示为源蛋白的注释值(例如5'UTR长度)。在另一个实例中，编码模块314通过在等位基因非相互作用变量w^i中包括指示变量来表示p^i的残基水平的源蛋白注释，即，如果肽p^i与螺旋基元重叠则等于1，否则是0，或者如果肽p^i完全包含在螺旋基元内则等于1。在另一个实例中，表示肽p^i中包含在螺旋基元注释内的残基的比例的特征可以包括在等位基因非相互作用变量w^i中。

[0395] 在一个实例中，编码模块314将人蛋白质组中蛋白质或同功型的类型表示为指示向量o^k，所述向量的长度等于人蛋白质组中蛋白质或同功型的数量，并且如果肽p^k来自蛋白质i，则相应元素$o^k{}_i$是1，否则是0。

[0396] 在一个实例中，编码模块314将肽p^i的源基因G＝基因(p^i)表示为具有L个可能类别的分类变量，其中L表示索引的源基因1、2、…、L的数目的上限。

[0397] 在一个实例中，编码模块314将肽p^i的组织类型、细胞类型、肿瘤类型或肿瘤组织学类型T＝组织(p^i)表示为具有M个可能类别的分类变量，其中M表示索引的类型1、2、…、M的数目的上限。组织的类型可以包括例如肺组织、心脏组织、肠组织、神经组织等。细胞类型可以包括树突细胞、巨噬细胞、CD4T细胞等。肿瘤类型可以包括肺腺癌、肺鳞状细胞癌、黑素瘤、非霍奇金淋巴瘤等。

[0398] 编码模块314还可以将有关肽p^i和相关MHC等位基因h的变量z^i的总体集合表示为行向量，其中等位基因相互作用变量x^i和等位基因非相互作用变量w^i的数字表示相继地串接。举例来说，编码模块314可以将$z_h{}^i$表示为等于[$x_h{}^i$ w^i]或[w_i $x_h{}^i$]的行向量。

[0399] IX.训练模块

[0400] 训练模块316构建一个或多个呈递模型，这些模型生成肽序列是否会被与这些肽序列相关的MHC等位基因呈递的可能性。确切地说，给定肽序列p^k和与肽序列p^k相关联的MHC等位基因集合a^k，每个呈递模型生成估计值u_k，指示肽序列p^k会被与一个或多个相关MHC等位基因a^k呈递的可能性。

[0401] IX.A.概述

[0402] 训练模块316基于由存储在165中的呈递信息产生的存储于存储器170中的训练数据集来构建一个或多个呈递模型。一般来说，不管呈递模型的具体类型如何，所有呈递模型都捕捉训练数据170中自变量与因变量之间的相关性以使损失函数减到最小。确切地说，损失函数($y_{i\in S}$,$u_{i\in S}$;θ)表示训练数据170中一个或多个数据实例S的因变量$y_{i\in S}$与由呈递模型生成的数据实例S的估计可能性$u_{i\in S}$值之间的偏差。在本说明书其余部分通篇所提到的一个

特定实施方式中，损失函数($y_{i\in S}, u_{i\in S}; \theta$)是由以下等式(1a)提供的负对数可能性函数：

[0403] $$\ell(y_{i\in S}, u_{i\in S}; \boldsymbol{\theta}) = \sum_{i\in S}(y_i \log u_i + (1-y_i)\log(1-u_i))。 \quad (1a)$$

[0404] 不过，实际上，可以使用另一损失函数。举例来说，当对质谱离子电流进行预测时，损失函数是由以下等式1b提供的均方损失：

[0405] $$\ell(y_{i\in S}, u_{i\in S}; \theta) = \sum_{i\in S}(\|y_i - u_i\|_2^2)。 \quad (1b)$$

[0406] 呈递模型可以是一种参数模型，其中一个或多个参数θ在数学上指示自变量与因变量之间的相关性。通常，使损失函数($y_{i\in S}, u_{i\in S}; \theta$)最小的参数型呈递模型的各种参数是通过基于梯度的数值优化算法，如批量梯度算法、随机梯度算法等来确定。或者，呈递模型可以是非参数模型，其中模型结构是由训练数据170决定并且并不严格基于固定参数集合。

[0407] IX.B.独立等位基因模型

[0408] 训练模块316可以在独立等位基因(per-allele)基础上构建呈递模型以预测肽的呈递可能性。在此情况下，训练模块316可以基于由表达单个MHC等位基因的细胞产生的训练数据170中的数据实例S训练呈递模型。

[0409] 在一个实施方式中，训练模块316通过下式使特定等位基因h对于肽p^k的估计呈递可能性u_k建模：

[0410] $$u_k^h = \Pr(\text{呈递的}\boldsymbol{p^k}; \text{MHC等位基因}\, h) = f\left(g_h\left(\boldsymbol{x_h^k}; \boldsymbol{\theta_h}\right)\right), \quad (2)$$

[0411] 其中肽序列x_h^k表示编码的有关肽p^k和相应MHC等位基因h的等位基因相互作用变量，f(•)是任何函数，并且为便于说明，在本文通篇称为变换函数。此外，g_h(•)是任何函数，为便于说明，在本文通篇称为相关性函数(dependency function)，并且基于所测定的MHC等位基因h的参数集合θ_h产生等位基因相互作用变量x_h^k的相关性分数。有关各MHC等位基因h的参数集合θ_h的值可以通过使关于θ_h的损失函数减到最小来测定，其中i是由表达单个MHC等位基因h的细胞所产生的训练数据170的子集S中的每个实例。

[0412] 相关性函数$g_h(x_h^k; \theta_h)$的输出值表示至少基于等位基因相互作用特征x_h^k，并且确切地说，基于肽p^k的肽序列中氨基酸的位置的针对MHC等位基因h的相关性分数，其指示MHC等位基因h将呈递相应新抗原。举例来说，如果MHC等位基因h可能呈递肽p^k，则MHC等位基因h的相关性分数可能具有较高值，而如果不可能呈递，则可能具有较低值。变换函数f(•)将输入，并且更确切地说，在此情形中将由$g_h(x_h^k; \theta_h)$生成的相关性分数变换成适当值以指示肽pk将由MHC等位基因呈递的可能性。

[0413] 在本说明书其余部分通篇提到的一个特定实施方式中，f(•)是对于适当域范围具有在[0,1]内的范围的函数。在一个实施例中，f(•)是由下式提供的expit函数：

[0414] $$f(z) = \frac{\exp(z)}{1+\exp(z)}。 \quad (4)$$

[0415] 又如，当域z的值等于或大于0时，f(•)也可以是由下式提供的双曲正切函数：

[0416] $f(z) = \tanh(z)$ (5)

[0417] 或者，当质谱离子电流的预测值超出范围[0,1]时，f(•)可以是任何函数，如恒等函数、指数函数、对数函数等。

[0418] 因此，可以通过将有关MHC等位基因h的相关性函数$g_h(\bullet)$应用于肽序列p^k的编码形式以产生相应相关性分数来产生肽序列p^k将由MHC等位基因h呈递的独立等位基因可能性。相关性分数可以由变换函数$f(\bullet)$变换以产生肽序列p^k将由MHC等位基因h呈递的独立等位基因可能性。

[0419] IX.B.1有关等位基因相互作用变量的相关性函数

[0420] 在本发明通篇提到的一个特定实施方式中，相关性函数$g_h(\bullet)$是由下式提供的仿射函数：

[0421] $$g_h(\boldsymbol{x}_h^i;\boldsymbol{\theta}_h)=\boldsymbol{x}_h^i\cdot\boldsymbol{\theta}_h。 \quad (6)$$

[0422] 所述函数将x_h^k中的每个等位基因相互作用变量与所测定的相关MHC等位基因h的参数集合θ_h中的相应参数线性地组合。

[0423] 在本说明书通篇提到的另一个特定实施方式中，相关性函数$g_h(\bullet)$是由下式提供的网络函数：

[0424] $$g_h(\boldsymbol{x}_h^i;\boldsymbol{\theta}_h)=NN_h(\boldsymbol{x}_h^i;\boldsymbol{\theta}_h)。 \quad (7)$$

[0425] 以具有分一层或多层布置的一系列节点的网络模型$NN_h(\bullet)$表示。一个节点可以通过连接而连接至其他节点，这些连接各自在参数集合θ_h中具有相关参数。在一个特定节点处的值可以表示为通过与所述特定节点相关联的激活函数所映射的相关参数加权的连接至所述特定节点的节点值的总和。由于呈递模型可以并入具有不同氨基酸序列长度的非线性和工艺数据，与仿射函数相比，网络模型是有利的。确切地说，通过非线性建模，网络模型可以捕捉在肽序列不同位置处的氨基酸之间的相互作用以及这一相互作用如何影响肽呈递。

[0426] 一般来说，网络模型$NN_h(\bullet)$可以被构造成前馈网络，如人工神经网络(ANN)、卷积神经网络(CNN)、深度神经网络(DNN)，和/或循环网络，如长短期记忆网络(LSTM)、双向循环网络、深度双向循环网络等。

[0427] 在本说明书其余部分通篇提到的一个实例中，$h=1,2,\cdots,m$中的每个MHC等位基因与独立网络模型相关联，并且$NN_h(\bullet)$表示来自与MHC等位基因h相关联的网络模型的输出。

[0428] 图5示出了与任意MHC等位基因$h=3$相关联的示例性网络模型$NN_3(\bullet)$。如图5中所示，关于MHC等位基因$h=3$的网络模型$NN_3(\bullet)$包括在层$l=1$处的三个输入节点、在层$l=2$处的四个节点、在层$l=3$处的两个节点和在层$l=4$处的一个输出节点。网络模型$NN_3(\bullet)$与十个参数集合$\theta_3(1)$、$\theta_3(2)$、…、$\theta_3(10)$相关。网络模型$NN_3(\bullet)$接收关于MHC等位基因$h=3$的三个等位基因相互作用变量$x_3^k(1)$、$x_3^k(2)$和$x_3^k(3)$的输入值(包括编码的多肽序列数据和所用任何其他训练数据的个别数据实例)并输出值$NN_3(x_3^k)$。网络函数还可以包括一个或多个网络模型，每个网络模型采用不同的等位基因相互作用变量作为输入。

[0429] 在另一个实施例中，鉴别的MHC等位基因$h=1,2,\cdots,m$与单个网络模型$NN_H(\bullet)$相关联，并且$NN_H(\bullet)$表示与MHC等位基因h相关的单个网络模型的一个或多个输出。在此类实例中，参数集合θ_h可以对应于所述单个网络模型的参数集合，并因此，参数集合θ_h可以是所有MHC等位基因共有的。

[0430] 图6A示出了MHC等位基因$h=1,2,\cdots,m$共有的示例性网络模型$NN_H(\bullet)$。如图6A中所示，网络模型$NN_H(\bullet)$包括m个输出节点，各自对应于MHC等位基因。网络模型$NN_H(\bullet)$接收

有关MHC等位基因h＝3的等位基因相互作用变量x_3^k并输出m值，包括对应于MHC等位基因h＝3的值$NN_3(x_3^k)$。

[0431] 在又另一实例中，单个网络模型$NN_H(\cdot)$可以是在给定MHC等位基因h的等位基因相互作用变量x_h^k和编码的蛋白质序列d_h情况下，输出相关性分数的网络模型。在此类实例中，参数集合θ_h也可以对应于所述单个网络模型的参数集合，并因此，参数集合θ_h可以是所有MHC等位基因共有的。因此，在此类实例中，$NN_h(\cdot)$可以表示在给定所述单个网络模型的输入$[x_h^k\ d_h]$情况下，所述单个网络模型$NN_H(\cdot)$的输出。由于训练数据中未知的MHC等位基因的肽呈递可能性只能通过鉴别其蛋白质序列进行预测，故此类网络模型是有利的。

[0432] 图6B示出了MHC等位基因共有的示例性网络模型$NN_H(\cdot)$。如图6B中所示，网络模型$NN_H(\cdot)$接收MHC等位基因h＝3的等位基因相互作用变量和蛋白质序列作为输入，并输出对应于MHC等位基因h＝3的相关性分数$NN_3(x_3^k)$。

[0433] 在又另一个实施例中，相关性函数$g_h(\cdot)$可以表示为：

[0434] $$g_h\left(x_h^k;\theta_h\right)=g'_h\left(x_h^k;\theta'_h\right)+\theta_h^0$$

[0435] 其中$g'_h(x_h^k;\theta'_h)$是具有参数集合θ'_h的仿射函数、网络函数等，其中有关MHC等位基因的等位基因相互作用变量的参数集合的偏差参数θ_h^0表示MHC等位基因h的基线呈递机率。

[0436] 在另一个实施方式中，偏差参数θ_h^0可以是MHC等位基因h的基因家族共有的。也就是说，MHC等位基因h的偏差参数θ_h^0可以等于$\theta_{基因(h)}^0$，其中基因(h)是MHC等位基因h的基因家族。举例来说，I类MHC等位基因HLA-A*02:01、HLA-A*02:02和HLA-A*02:03可以指定给“HLA-A”基因家族，并且这些MHC等位基因各自的偏差参数θ_h^0可以是共有的。又如，II类MHC等位基因HLA-DRB1:10:01、HLA-DRB1:11:01和HLA-DRB3:01:01可以指定给“HLA-DRB”基因家族，并且这些MHC等位基因各自的偏差参数θ_h^0可以是共有的。

[0437] 再回到等式(2)，例如，在使用仿射相关性函数$g_h(\cdot)$鉴别的m＝4种不同的MHC等位基因当中，肽p^k将由MHC等位基因h＝3呈递的可能性可以由下式得到：

[0438] $$u_k^3=f\left(x_3^k\cdot\theta_3\right),$$

[0439] 其中x_3^k是鉴别的MHC等位基因h＝3的等位基因相互作用变量，并且θ_3是通过损失函数最小化测定的MHC等位基因h＝3的参数集合。

[0440] 又如，在使用独立网络变换函数$g_h(\cdot)$鉴别的m＝4种不同的MHC等位基因当中，肽p^k将由MHC等位基因h＝3呈递的可能性可以由下式得到：

[0441] $$u_k^3=f\left(NN_3(x_3^k;\theta_3)\right),$$

[0442] 其中x_3^k是鉴别的MHC等位基因h＝3的等位基因相互作用变量，并且θ_3是测定的与MHC等位基因h＝3相关联的网络模型$NN_3(\cdot)$的参数集合。

[0443] 图7示出了使用示例性网络模型$NN_3(\cdot)$生成与MHC等位基因h＝3相关联的肽p^k的呈递可能性。如图7中所示，网络模型$NN_3(\cdot)$接收有关MHC等位基因h＝3的等位基因相互作用变量x_3^k并生成输出$NN_3(x_3^k)$。所述输出由函数$f(\cdot)$映射以产生估计的呈递可能性u_k。

[0444] IX.B.2.具有等位基因非相互作用变量的独立等位基因

[0445] 在一个实施方式中，训练模块316并入等位基因非相互作用变量并通过下式使肽p^k的估计呈递可能性u_k建模：

[0446] $$u_k^h = \Pr(\text{呈递的 } \boldsymbol{p^k}) = f\left(g_w(\boldsymbol{w^k};\boldsymbol{\theta_w}) + g_h(\boldsymbol{x_h^i};\boldsymbol{\theta_h})\right), \quad (8)$$

[0447] 其中w^k表示肽p^k的编码的等位基因非相互作用变量，$g_w(\cdot)$是基于测定的等位基因非相互作用变量的参数集合θ_w的等位基因非相互作用变量w^k的函数。确切地说，有关各MHC等位基因h的参数集合θ_h和有关等位基因非相互作用变量的参数集合θ_w的值可以通过使关于θ_h和θ_w的损失函数减到最小来测定，其中i是由表达单个MHC等位基因的细胞所产生的训练数据170的子集S中的每个实例。

[0448] 相关性函数$g_w(w^k;\theta_w)$的输出表示基于等位基因非相互作用变量的影响的等位基因非相互作用变量的相关性分数，其指示肽p^k是否会由一个或多个MHC等位基因呈递。举例来说，如果肽p^k与已知会积极地影响肽p^k的呈递的C末端侧接序列相关，则等位基因非相互作用变量的相关性分数可能具有较高值，并且如果肽p^k与已知会不利地影响肽p^k的呈递的C末端侧接序列相关，则可能具有较低值。

[0449] 根据等式(8)，可以通过将有关MHC等位基因h的函数$g_h(\cdot)$应用于肽序列p^k的编码形式以产生等位基因相互作用变量的相应相关性分数来产生肽序列p^k将由MHC等位基因h呈递的独立等位基因可能性。有关等位基因非相互作用变量的函数$g_w(\cdot)$也应用于等位基因非相互作用变量的编码形式以产生等位基因非相互作用变量的相关性分数。将两个分数合并，并通过变换函数$f(\cdot)$变换所述合并的分数以产生肽序列p^k将由MHC等位基因h呈递的独立等位基因可能性。

[0450] 或者，训练模块316可以通过将等位基因非相互作用变量w^k添加至等式(2)中的等位基因非相互作用变量x_h^k中，来将等位基因非相互作用变量w^k包括在预测值中。因此，呈递可能性可以由下式得到：

[0451] $$u_k^h = \Pr(\text{呈递的}\boldsymbol{p^k};\text{等位基因}h) = f\left(g_h([\boldsymbol{x_h^k}\ \boldsymbol{w^k}];\boldsymbol{\theta_h})\right)。 \quad (9)$$

[0452] IX.B.3有关等位基因非相互作用变量的相关性函数

[0453] 与有关等位基因相互作用变量的相关性函数$g_h(\cdot)$类似，有关等位基因非相互作用变量的相关性函数$g_w(\cdot)$可以是仿射函数或网络函数，其中独立网络模型与等位基因非相互作用变量w^k相关联。

[0454] 确切地说，相关性函数$g_w(\cdot)$是由下式提供的仿射函数：

[0455] $g_w(w^k;\theta_w) = w^k \cdot \theta_w$。

[0456] 所述函数将等位基因非相互作用变量w^k与参数集合θ_w中的相应参数线性地组合。

[0457] 相关性函数$g_w(\cdot)$还可以是由下式提供的网络函数：

[0458] $g_h(w^k;\theta_w) = NN_w(w^k;\theta_w)$。

[0459] 所述函数是由具有参数集合θ_w中的相关参数的网络模型$NN_w(\cdot)$表示。网络函数可能还包括一个或多个网络模型，每个网络模型采用不同的等位基因非相互作用变量作为输入。

[0460] 在另一个实施例中，有关等位基因非相互作用变量的相关性函数$g_w(\cdot)$可以由下式提供：

[0461] $$g_w(\boldsymbol{w^k};\boldsymbol{\theta_w}) = g'_w(\boldsymbol{w^k};\boldsymbol{\theta'_w}) + h(m^k;\theta_w^m), \quad (10)$$

[0462] 其中$g'_w(w^k;\theta'_w)$是仿射函数，具有等位基因非相互作用参数集合θ'_w的网络函数

等，m^k是肽p^k的mRNA定量测量值，h(•)是变换所述定量测量值的函数，并且θ_w^m是有关等位基因非相互作用变量的参数集合中的一个参数，所述参数与mRNA定量测量值组合以生成有关mRNA定量测量值的相关性分数。在本说明书其余部分通篇所提到的一个特定实施方案中，h(•)是对数函数，不过实际上，h(•)可以是多种不同函数中的任一种。

[0463] 在又另一个实例中，有关等位基因非相互作用变量的相关性函数$g_w(\bullet)$可以由下式提供：

[0464] $$g_w(\boldsymbol{w}^k;\boldsymbol{\theta}_w)=g'_w(\boldsymbol{w}^k;\boldsymbol{\theta}'_w)+\boldsymbol{\theta}_w^o\cdot\boldsymbol{o}^k, \qquad (11)$$

[0465] 其中$g'_w(w^k;\theta'_w)$是仿射函数、具有等位基因非相互作用参数集合θ'_w的网络函数等，θ'_w是部分VII.C.2中描述的表示人蛋白质组中有关肽p^k的蛋白质和同功型的指示向量，并且θ_w^o是有关等位基因非相互作用变量的参数集合中的参数集合，其与指示向量组合。在一种变化形式中，当o^k的维度和参数集合θ_w^o明显较高时，可以在测定参数值时将参数正则项，诸如$\lambda\cdot\|\boldsymbol{\theta}_w^o\|$添加至损失函数中，其中||•||表示L1范数、L2范数、组合等。超参数λ的最佳值可以通过适当方法测定。

[0466] 在又另一个实例中，有关等位基因非相互作用变量的相关性函数$g_w(\bullet)$可以由下式提供：

[0467] $$g_w(\boldsymbol{w}^k;\boldsymbol{\theta}_w)=g'_w(\boldsymbol{w}^k;\boldsymbol{\theta}'_w)+\sum_{l=1}^{L}\mathbb{1}\left(\text{基因}(\boldsymbol{p}^k=l)\right)\cdot\boldsymbol{\theta}_w^l, \qquad (12)$$

[0468] 其中$g'_w(w^k;\theta'_w)$是仿射函数、具有等位基因非相互作用参数集合θ'_w的网络函数等，(基因(p^k=1))是指示函数，如上文对于等位基因非相互作用变量所述，如果肽p^k来自源基因1，则其等于1，并且p^k是指示源基因1的“抗原性”的参数。在一种变化形式中，当L显著较高并且因此参数$\theta_w^{l=1,2,\cdots,L}$数量也显著较高时，可以在测定参数值时将参数正则项，诸如$\lambda\cdot\|\boldsymbol{\theta}_w^l\|$添加至损失函数中，其中||•||表示L1范数、L2范数、组合等。超参数λ的最佳值可以通过适当方法测定。

[0469] 在又另一个实例中，有关等位基因非相互作用变量的相关性函数$g_w(\bullet)$可以由下式提供：

[0470] $$g_w(\boldsymbol{w}^k;\boldsymbol{\theta}_w)=g'_w(\boldsymbol{w}^k;\boldsymbol{\theta}'_w)+\sum_{m=1}^{M}\sum_{l=1}^{L}\mathbb{1}(\text{基因}(\boldsymbol{p}^k)=l,\text{组织}(\boldsymbol{p}^k)=m)\cdot\boldsymbol{\theta}_w^{lm}, \qquad (12b)$$

[0471] 其中$g'_w(w^k;\theta'_w)$是仿射函数、具有等位基因非相互作用参数集合θ'_w的网络函数等，(基因(p^k)=1，组织(p^k)=m)是指示函数，如上文对于等位基因非相互作用变量所述，如果肽p^k来自源基因1并且如果肽p^k来自组织类型m，则其等于1，并且θ_w^{lm}是指示源基因1和组织类型m的组合的抗原性的参数。确切地说，组织类型m的基因1的抗原性可以表示在控制RNA表达和肽序列背景之后，组织类型m的细胞呈递来自基因1的肽的残余倾向。

[0472] 在一种变化形式中，当L或M显著较高并且因此参数$\theta_w^{lm=1,2,\cdots,LM}$数量也显著较高时，可以在测定参数值时将参数正则项，诸如$\lambda\cdot\|\boldsymbol{\theta}_w^{lm}\|$添加至损失函数中，其中||•||表示L1范数、L2范数、组合等。超参数λ的最佳值可以通过适当方法测定。在另一种变化形式

中，可以在测定参数值时将参数正则项添加至损失函数中，使得相同源基因的系数不会在组织类型之间有显著差异。例如，惩罚项诸如：

[0473] $$\lambda \cdot \sum_{l=1}^{L} \sqrt{\sum_{m=1}^{M} \left(\theta_w^{lm} - \overline{\theta_w^l}\right)^2}$$

[0474] 可以惩罚损失函数中不同组织类型之间抗原性的标准偏差，其中$\overline{\theta_w^l}$是源基因1的组织类型之间的平均抗原性。

[0475] 实际上，等式(10)、(11)、(12a)和(12b)中的任一个的附加项可以组合以产生等位基因非相互作用变量的相关性函数$g_w(\bullet)$。例如，可以将等式(10)中表示mRNA定量测量的项$h(\bullet)_i$和等式(12)中表示源基因抗原性的项与任何其他仿射或网络函数一起相加，以生成等位基因非相互作用变量的相关性函数。

[0476] 再回到等式(8)，例如，在使用仿射变换函数$g_h(\bullet)$、$g_w(\bullet)$鉴别的m＝4种不同的MHC等位基因当中，肽p^k将由MHC等位基因h＝3呈递的可能性可以由下式产生：

[0477] $$u_k^3 = f\left(\boldsymbol{w}^k \cdot \boldsymbol{\theta}_w + \boldsymbol{x}_3^k \cdot \boldsymbol{\theta}_3\right),$$

[0478] 其中w^k是所鉴别的肽p^k的等位基因非相互作用变量，并且θ_w是测定的等位基因非相互作用变量的参数的集合。

[0479] 又如，在使用网络变换函数$g_h(\bullet)$、$g_w(\bullet)$鉴别的m＝4种不同的MHC等位基因当中，肽p^k将由MHC等位基因h＝3呈递的可能性可以由下式得到：

[0480] $$u_k^3 = f\left(NN_w\left(\boldsymbol{w}^k;\boldsymbol{\theta}_w\right) + NN_3\left(\boldsymbol{x}_3^k;\boldsymbol{\theta}_3\right)\right)$$

[0481] 其中w^k是所鉴别的肽p^k的等位基因相互作用变量，并且θ_w是测定的等位基因非相互作用变量的参数的集合。

[0482] 图8示出了使用示例性网络模型$NN_3(\bullet)$和$NN_w(\bullet)$生成与MHC等位基因h＝3相关联的肽p^k的呈递可能性。如图8中所示，网络模型$NN_3(\bullet)$接收有关MHC等位基因h＝3的等位基因相互作用变量$x_3{}^k$并生成输出$NN_3(x_3{}^k)$。网络模型$NN_w(\bullet)$接收有关肽p^k的等位基因非相互作用变量w^k并生成输出$NN_w(w^k)$。将输出合并，并由函数$f(\bullet)$映射以产生估计的呈递可能性u_k。

[0483] IX.C.多等位基因模型

[0484] 训练模块316还可以在存在两个或更多个MHC等位基因的多等位基因环境中构建呈递模型以预测肽的呈递可能性。在此情况下，训练模块316可以基于由表达单个MHC等位基因的细胞、表达多个MHC等位基因的细胞或其组合产生的训练数据170中的数据实例S训练呈递模型。

[0485] IX.C.1.实施例1：独立等位基因模型的最大化

[0486] 在一个实施方式中，训练模块316使与多个MHC等位基因H的集合相关联的肽p^k的估计呈递可能性u_k随基于表达单等位基因的细胞所测定的集合H中每个MHC等位基因h的呈递可能性$u_k{}^{h\in H}$的变化建模，如上文结合等式(2)-(11)所描述。确切地说，呈递可能性u_k可以是$u_k{}^{h\in H}$的任何函数。在一个实施方式中，如等式(12)中所示，所述函数是最大值函数，并且

呈递可能性u_k可以测定为集合H中每个MHC等位基因h的呈递可能性最大值。

[0487] $u_k = \Pr(\text{呈递的}\boldsymbol{p^k};\text{等位基因}H) = \text{最大值}(u_k^{h\in H})$。

[0488] IX.C.2.实施例2.1:和的函数(Funciton-of-Sums)模型

[0489] 在一个实施方式中,训练模块316通过下式使肽p^k的估计呈递可能性u_k建模:

[0490] $$u_k = \Pr(\text{呈递的}\boldsymbol{p^k}) = f\left(\sum_{h=1}^{m} a_h^k \cdot g_h(\boldsymbol{x_h^k};\boldsymbol{\theta_h})\right), \qquad (13)$$

[0491] 其中元素a_h^k对于与肽序列p^k相关的多个MHC等位基因H是1,并且x_h^k表示编码的有关肽p^k和相应MHC等位基因的等位基因相互作用变量。有关各MHC等位基因h的参数集合θ_h的值可以通过使关于θ_h的损失函数减到最小来测定,其中i是由表达单个MHC等位基因的细胞和/或表达多个MHC等位基因的细胞所产生的训练数据170的子集S中的每个实例。相关性函数g_h可以呈以上VIII.B.1部分中介绍的相关性函数g_h中的任一种的形式。

[0492] 根据等式(13),可以通过将相关性函数$g_h(\bullet)$应用于有关MHC等位基因H中的每一个的肽序列p^k的编码形式以产生等位基因相互作用变量的相应分数来产生肽序列p^k将由一个或多个MHC等位基因h呈递的呈递可能性。将每个MHC等位基因h的分数合并,并通过变换函数$f(\bullet)$变换以产生肽序列p^k将由MHC等位基因集合H呈递的呈递可能性。

[0493] 等式(13)的呈递模型与等式(2)的独立等位基因模型的不同之处在于,每个肽p^k的相关等位基因的数量可以大于1。换句话说,对于与肽序列p^k相关的多个MHC等位基因H,$a_h^k a_h^k$中超过一个元素的值可以是1。

[0494] 例如,在使用仿射变换函数$g_h(\bullet)$鉴别的m=4种不同的MHC等位基因当中,肽p^k将由MHC等位基因h=2、h=3呈递的可能性可以由下式得到:

[0495] $$u_k = f\left(\boldsymbol{x_2^k}\cdot\boldsymbol{\theta_2} + \boldsymbol{x_3^k}\cdot\boldsymbol{\theta_3}\right),$$

[0496] 其中x_2^k、x_3^k是鉴别的MHC等位基因h=2、h=3的等位基因相互作用变量,并且θ_2、θ_3是测定的MHC等位基因h=2、h=3的参数的集合。

[0497] 又如,在使用网络变换函数$g_h(\bullet)$、$g_w(\bullet)$鉴别的m=4种不同的MHC等位基因当中,肽p^k将由MHC等位基因h=2、h=3呈递的可能性可以由下式得到:

[0498] $$\boldsymbol{u_k} = f\left(NN_2(\boldsymbol{x_2^k};\boldsymbol{\theta_2}) + NN_3(\boldsymbol{x_3^k};\boldsymbol{\theta_3})\right),$$

[0499] 其中$NN_2(\bullet)$、$NN_3(\bullet)$是鉴别的MHC等位基因h=2、h=3的网络模型,并且θ_2、θ_3是测定的MHC等位基因h=2、h=3的参数的集合。

[0500] 图9示出了使用示例性网络模型$NN_2(\bullet)$和$NN_3(\bullet)$生成与MHC等位基因h=2、h=3相关联的肽p^k的呈递可能性。如图9中所示,网络模型$NN_2(\bullet)$接收有关MHC等位基因h=2的等位基因相互作用变量x_2^k并生成输出$NN_2(x_2^k)$,并且网络模型$NN_3(\bullet)$接收有关MHC等位基因h=3的等位基因相互作用变量x_3^k并生成输出$NN_3(x_3^k)$。将输出合并,并由函数$f(\bullet)$映射以产生估计的呈递可能性u_k。

[0501] IX.C.3.实施例2.2:利用等位基因非相互作用变量的和的函数模型

[0502] 在一个实施方式中,训练模块316并入等位基因非相互作用变量并通过下式使肽p^k的估计呈递可能性u_k建模:

[0503] $$u_k = \Pr(\text{呈递的}\boldsymbol{p^k}) = f\left(g_w(\boldsymbol{w^k};\boldsymbol{\theta_w}) + \sum_{h=1}^{m} a_h^k \cdot g_h(\boldsymbol{x_h^k};\boldsymbol{\theta_h})\right), \qquad (14)$$

[0504] 其中w^k表示编码的有关肽p^k的等位基因非相互作用变量。确切地说，有关各MHC等位基因h的参数集合θ_h和有关等位基因非相互作用变量的参数集合θ_w的值可以通过使关于θ_h和θ_w的损失函数减到最小来测定，其中i是由表达单个MHC等位基因的细胞和/或表达多个MHC等位基因的细胞所产生的训练数据170的子集S中的每个实例。相关性函数g_w可以呈以上VIII.B.3部分中介绍的相关性函数g_w中的任一种的形式。

[0505] 因此，根据等式(14)，可以通过将函数$g_h(\bullet)$应用于有关MHC等位基因H中的每一个的肽序列p^k的编码形式以产生有关每个MHC等位基因h的等位基因相互作用变量的相应相关性分数来产生肽序列p^k将由一个或多个MHC等位基因H呈递的呈递可能性。有关等位基因非相互作用变量的函数$g_w(\bullet)$也应用于等位基因非相互作用变量的编码形式以产生等位基因非相互作用变量的相关性分数。将分数合并，并通过变换函数$f(\bullet)$变换所述合并的分数以产生肽序列p^k将由MHC等位基因H呈递的呈递可能性。

[0506] 在等式(14)的呈递模型中，每个肽p^k的相关等位基因的数量可以大于1。换句话说，对于与肽序列pk相关的多个MHC等位基因H，ahkahk中超过一个元素的值可以是1。

[0507] 例如，在使用仿射变换函数$g_h(\bullet)$、$g_w(\bullet)$鉴别的m＝4种不同的MHC等位基因当中，肽p^k将由MHC等位基因h＝2、h＝3呈递的可能性可以由下式得到：

[0508] $$u_k = f\left(\boldsymbol{w^k}\cdot\boldsymbol{\theta_w} + \boldsymbol{x_2^k}\cdot\boldsymbol{\theta_2} + \boldsymbol{x_3^k}\cdot\boldsymbol{\theta_3}\right),$$

[0509] 其中w^k是所鉴别的肽p^k的等位基因非相互作用变量，并且θ_w是测定的等位基因非相互作用变量的参数的集合。

[0510] 又如，在使用网络变换函数$g_h(\bullet)$、$g_w(\bullet)$鉴别的m＝4种不同的MHC等位基因当中，肽p^k将由MHC等位基因h＝2、h＝3呈递的可能性可以由下式得到：

[0511] $$u_k = f\left(NN_w(\boldsymbol{w^k};\boldsymbol{\theta_w}) + NN_2(\boldsymbol{x_2^k};\boldsymbol{\theta_2}) + NN_3(\boldsymbol{x_3^k};\boldsymbol{\theta_3})\right)$$

[0512] 其中w^k是所鉴别的肽p^k的等位基因相互作用变量，并且θ_w是测定的等位基因非相互作用变量的参数的集合。

[0513] 图10示出了使用示例性网络模型$NN_2(\bullet)$、$NN_3(\bullet)$和$NN_w(\bullet)$生成与MHC等位基因h＝2、h＝3相关联的肽p^k的呈递可能性。如图10中所示，网络模型$NN_2(\bullet)$接收有关MHC等位基因h＝2的等位基因相互作用变量x_2^k并生成输出$NN_2(x_2^k)$。网络模型$NN_3(\bullet)$接收有关MHC等位基因h＝3的等位基因相互作用变量x_3^k并生成输出$NN_3(x_3^k)$。网络模型$NN_w(\bullet)$接收有关肽p^k的等位基因非相互作用变量w^k并生成输出$NN_w(w^k)$。将输出合并，并由函数$f(\bullet)$映射以产生估计的呈递可能性u_k。

[0514] 或者，训练模块316可以通过将等位基因非相互作用变量w^k添加至等式(15)中的等位基因非相互作用变量x_h^k中，来将等位基因非相互作用变量w^k包括在预测值中。因此，呈递可能性可以由下式得到：

[0515] $$u_k = \Pr(\text{呈递的}\boldsymbol{p^k}) = f\left(\sum_{h=1}^{m} a_h^k \cdot g_h([\boldsymbol{x_h^k}\,\boldsymbol{w^k}];\boldsymbol{\theta_h})\right)。 \qquad (15)$$

[0516] IX.C.4.实施例3.1：使用隐式独立等位基因可能性的模型

[0517] 在另一个实施方式中，训练模块316通过下式使肽p^k的估计呈递可能性u_k建模：

[0518] $$u_k = \Pr(\text{呈递的}\boldsymbol{p^k}) = r\left(s\left(\boldsymbol{v} = \left[a_1^k \cdot u'^1_k(\boldsymbol{\theta}) \ldots a_m^k \cdot u'^m_k(\boldsymbol{\theta})\right]\right)\right), \quad (16)$$

[0519] 其中元素a_h^k对于与肽序列p^k相关联的多个MHC等位基因$h \in H$是1，u'^h_k是MHC等位基因h的隐式独立等位基因呈递可能性，向量v是其中元素v_h对应于$a_h^k \cdot u'^h_k$的向量，s(•)是映射元素v的函数，并且r(•)是限幅函数(clipping function)，其将输入值削减至给定范围中。如以下更详细地描述，s(•)可以是求和函数或二阶函数，但应理解在其他实施方案中，s(•)可以是任何函数，如最大值函数。有关隐式独立等位基因可能性的参数集合θ的值可以通过使关于θ的损失函数减到最小来测定，其中i是由表达单个MHC等位基因的细胞和/或表达多个MHC等位基因的细胞所产生的训练数据170的子集S中的每个实例。

[0520] 使等式(17)的呈递模型中的呈递可能性随各自对应于肽p^k将由个别MHC等位基因h呈递的可能性的隐式独立等位基因呈递可能性u'^h_k的变化建模。隐式独立等位基因可能性与VIII.B部分的独立等位基因呈递可能性的不同之处在于，有关隐式独立等位基因可能性的参数可以从多等位基因环境习得，其中除单等位基因环境外，呈递肽与相应MHC等位基因之间的直接关联也是未知的。因此，在多等位基因环境中，呈递模型不仅可以估计肽p^k是否会由作为整体的MHC等位基因集合H呈递，而且还可以提供指示最可能呈递肽p^k的MHC等位基因h的个别可能性$u'^{h\in H}_k$。其优势在于，呈递模型可以在无有关表达单MHC等位基因的细胞的训练数据存在下产生隐式可能性。

[0521] 在本说明书其余部分通篇提到的一个特定实施方式中，r(•)是具有范围[0,1]的函数。举例来说，r(•)可以是限幅函数：

[0522] r(z)＝最小值(最大值(z,0),1)，

[0523] 其中选择z与1之间的最小值作为呈递可能性u_k。在另一个实施方式中，当域z的值等于或大于0时，r(•)也可以是由下式提供的双曲正切函数：

[0524] r(z)＝tanh(z)。

[0525] IX.C.5.实施例3.2:函数的和(Sum-of-Functions)模型

[0526] 在一个特定实施方式中，s(•)是求和函数，并且呈递可能性是通过对隐式独立等位基因呈递可能性求和得到：

[0527] $$u_k = \Pr(\text{呈递的}\boldsymbol{p^k}) = r\left(\sum_{h=1}^{m} a_h^k \cdot u'^h_k(\boldsymbol{\theta})\right)。 \quad (17)$$

[0528] 在一个实施方式中，MHC等位基因h的隐式独立等位基因呈递可能性是由下式得到：

[0529] $$u'^h_k = f\left(g_h\left(\boldsymbol{x_h^k}; \boldsymbol{\theta_h}\right)\right), \quad (18)$$

[0530] 由此通过下式估计出呈递可能性：

[0531] $$u_k = \Pr(\text{呈递的}\boldsymbol{p^k}) = r\left(\sum_{h=1}^{m} a_h^k \cdot f\left(g_h\left(\boldsymbol{x_h^k}; \boldsymbol{\theta_h}\right)\right)\right)。 \quad (19)$$

[0532] 根据等式(19)，可以通过将函数g_h(•)应用于有关MHC等位基因H中的每一个的肽序列p^k的编码形式以产生等位基因相互作用变量的相应相关性分数来产生肽序列p^k将由一

个或多个MHC等位基因H呈递的呈递可能性。每个相关性分数都先通过函数f(•)变换以产生隐式独立等位基因呈递可能性$u'_k{}^h$。将独立等位基因可能性$u'_k{}^h$合并，并且可以将限幅函数应用于合并的可能性以将值削减至范围[0,1]中以产生肽序列p^k将由MHC等位基因集合H呈递的呈递可能性。相关性函数g_h可以呈以上VIII.B.1部分中介绍的相关性函数g_h中的任一种的形式。

[0533] 例如，在使用仿射变换函数$g_h(\bullet)$鉴别的m＝4种不同的MHC等位基因当中，肽p^k将由MHC等位基因h＝2、h＝3呈递的可能性可以由下式得到：

[0534] $$u_k = r\left(f\left(\boldsymbol{x}_2^k \cdot \boldsymbol{\theta}_2\right) + f\left(\boldsymbol{x}_3^k \cdot \boldsymbol{\theta}_3\right)\right),$$

[0535] 其中$x_2{}^k$、$x_3{}^k$是鉴别的MHC等位基因h＝2、h＝3的等位基因相互作用变量，并且θ_2、θ_3是测定的MHC等位基因h＝2、h＝3的参数的集合。

[0536] 又如，在使用网络变换函数$g_h(\bullet)$、$g_w(\bullet)$鉴别的m＝4种不同的MHC等位基因当中，肽p^k将由MHC等位基因h＝2、h＝3呈递的可能性可以由下式得到：

[0537] $$u_k = r\left(f\left(NN_2(\boldsymbol{x}_2^k;\boldsymbol{\theta}_2)\right) + f\left(NN_3(\boldsymbol{x}_3^k;\boldsymbol{\theta}_3)\right)\right),$$

[0538] 其中$NN_2(\bullet)$、$NN_3(\bullet)$是鉴别的MHC等位基因h＝2、h＝3的网络模型，并且θ_2、θ_3是测定的MHC等位基因h＝2、h＝3的参数的集合。

[0539] 图11示出了使用示例性网络模型$NN_2(\bullet)$和$NN_3(\bullet)$生成与MHC等位基因h＝2、h＝3相关联的肽p^k的呈递可能性。如图9中所示，网络模型$NN_2(\bullet)$接收有关MHC等位基因h＝2的等位基因相互作用变量$x_2{}^k$并生成输出$NN_2(x_2{}^k)$，并且网络模型$NN_3(\bullet)$接收有关MHC等位基因h＝3的等位基因相互作用变量$x_3{}^k$并生成输出$NN_3(x_3{}^k)$。每个输出由函数f(•)映射以产生估计的呈递可能性u_k。

[0540] 在另一个实施方式中，当预测质谱离子电流的对数时，r(•)是对数函数并且f(•)是指数函数。

[0541] IX.C.6.实施例3.3:利用等位基因非相互作用变量的函数的和模型

[0542] 在一个实施方式中，MHC等位基因h的隐式独立等位基因呈递可能性是由下式得到：

[0543] $$u'_k{}^h = f\left(g_h\left(\boldsymbol{x}_h^k;\boldsymbol{\theta}_h\right) + g_w\left(\boldsymbol{w}^k;\boldsymbol{\theta}_w\right)\right), \qquad (20)$$

[0544] 由此通过下式产生呈递可能性：

[0545] $$u_k = \Pr\left(\text{呈递的}\boldsymbol{p}^k\right) = r\left(\sum_{h=1}^{m} a_h^k \cdot f\left(g_w\left(\boldsymbol{w}^k;\boldsymbol{\theta}_w\right) + g_h\left(\boldsymbol{x}_h^k;\boldsymbol{\theta}_h\right)\right)\right), \qquad (21)$$

[0546] 以并入等位基因非相互作用变量对肽呈递的影响。

[0547] 根据等式(21)，可以通过将函数$g_h(\bullet)$应用于有关MHC等位基因H中的每一个的肽序列p^k的编码形式以产生有关每个MHC等位基因h的等位基因相互作用变量的相应相关性分数来产生肽序列p^k将由一个或多个MHC等位基因H呈递的呈递可能性。有关等位基因非相互作用变量的函数$g_w(\bullet)$也应用于等位基因非相互作用变量的编码形式以产生等位基因非相互作用变量的相关性分数。将等位基因非相互作用变量的分数与等位基因相互作用变量的各个相关性分数合并。每个合并的分数都通过函数f(•)变换以产生隐式独立等位基因呈递可能性。将隐式可能性合并，并且可以将限幅函数应用于合并的输出以将值削减至

范围[0,1]中以产生肽序列p^k将由MHC等位基因集合H呈递的呈递可能性。相关性函数g_w可以呈以上VIII.B.3部分中介绍的相关性函数g_w中的任一种的形式。

[0548] 例如,在使用仿射变换函数$g_h(\bullet)$、$g_w(\bullet)$鉴别的m=4种不同的MHC等位基因当中,肽p^k将由MHC等位基因h=2、h=3呈递的可能性可以由下式得到:

[0549] $$u_k = r\left(f\left(\boldsymbol{w}^k \cdot \boldsymbol{\theta}_w + \boldsymbol{x}_2^k \cdot \boldsymbol{\theta}_2\right) + f\left(\boldsymbol{w}^k \cdot \boldsymbol{\theta}_w + \boldsymbol{x}_3^k \cdot \boldsymbol{\theta}_3\right)\right),$$

[0550] 其中w^k是所鉴别的肽p^k的等位基因非相互作用变量,并且θ_w是测定的等位基因非相互作用变量的参数的集合。

[0551] 又如,在使用网络变换函数gh(•)、gw(•)鉴别的m=4种不同的MHC等位基因当中,肽pk将由MHC等位基因h=2、h=3呈递的可能性可以由下式得到:

[0552] $$u_k = r\left(f\left(NN_w\left(\boldsymbol{w}^k;\boldsymbol{\theta}_w\right) + NN_2\left(\boldsymbol{x}_2^k;\boldsymbol{\theta}_2\right)\right) + f\left(NN_w\left(\boldsymbol{w}^k;\boldsymbol{\theta}_w\right) + NN_3\left(\boldsymbol{x}_3^k;\boldsymbol{\theta}_3\right)\right)\right)$$

[0553] 其中w^k是所鉴别的肽p^k的等位基因相互作用变量,并且θ_w是测定的等位基因非相互作用变量的参数的集合。

[0554] 图12示出了使用示例性网络模型$NN_2(\bullet)$、$NN_3(\bullet)$和$NN_w(\bullet)$生成与MHC等位基因h=2、h=3相关联的肽p^k的呈递可能性。如图12中所示,网络模型$NN_2(\bullet)$接收有关MHC等位基因h=2的等位基因相互作用变量x_2^k并生成输出$NN_2(x_2^k)$。网络模型$NN_w(\bullet)$接收有关肽p^k的等位基因非相互作用变量w^k并生成输出$NN_w(w^k)$。将输出合并,并且通过函数f(•)映射。网络模型$NN_3(\bullet)$接收有关MHC等位基因h=3的等位基因相互作用变量x_3^k并生成输出$NN_3(x_3^k)$,再次将所述输出与同一网络模型$NN_w(\bullet)$的输出$NN_w(w^k)$合并,并且通过函数f(•)映射。将两个输出合并以产生估计的呈递可能性u_k。

[0555] 在另一个实施方式中,MHC等位基因h的隐式独立等位基因呈递可能性由下式得到:

[0556] $$u_k'^h = f\left(g_h\left([\boldsymbol{x}_h^k\, \boldsymbol{w}^k];\boldsymbol{\theta}_h\right)\right)。 \qquad (22)$$

[0557] 由此通过下式产生呈递可能性:

[0558] $$u_k = \Pr(\text{呈递的}\boldsymbol{p}^k) = r\left(\sum_{h=1}^{m} a_h^k \cdot f\left(g_h\left([\boldsymbol{x}_h^k\, \boldsymbol{w}^k];\boldsymbol{\theta}_h\right)\right)\right)。$$

[0559] IX.C.7.实施例4:二阶模型

[0560] 在一个实施方式中,s(•)是二阶函数,并且肽p^k的估计呈递可能性u_k是由下式得到:

[0561] $$u_k = \Pr(\text{呈递的}\boldsymbol{p}^k) = \sum_{h=1}^{m} a_h^k \cdot u_k'^h(\boldsymbol{\theta}) - \sum_{h=1}^{m}\sum_{j<h} a_h^k \cdot a_j^k \cdot u_k'^h(\boldsymbol{\theta}) \cdot u_k'^j(\boldsymbol{\theta}) \qquad (23)$$

[0562] 其中元素$u'_k{}^h$是MHC等位基因h的隐式独立等位基因可能性。有关隐式独立等位基因可能性的参数集合θ的值可以通过使关于θ的损失函数减到最小来测定,其中i是由表达单个MHC等位基因的细胞和/或表达多个MHC等位基因的细胞所产生的训练数据170的子集S中的每个实例。隐式独立等位基因呈递可能性可以呈以上描述的等式(18)、(20)和(22)中所示的任何形式。

[0563] 在一方面,等式(23)的模型可以暗示存在肽p^k将同时由两个MHC等位基因呈递的

可能，其中两个HLA等位基因的呈递在统计学上是独立的。

[0564] 根据等式(23)，肽序列p^k将由一个或多个MHC等位基因H呈递的呈递可能性可以通过组合隐式独立等位基因呈递可能性并自总和中减去每对MHC等位基因将同时呈递肽p^k的可能性以产生肽序列p^k将由MHC等位基因H呈递的呈递可能性来产生。

[0565] 例如，在使用仿射变换函数$g_h(\bullet)$鉴别的m＝4种不同的HLA等位基因当中，肽p^k将由HLA等位基因h＝2、h＝3呈递的可能性可以由下式得到：

[0566] $$u_k = f(\boldsymbol{x}_2^k \cdot \boldsymbol{\theta}_2) + f(\boldsymbol{x}_3^k \cdot \boldsymbol{\theta}_3) - f(\boldsymbol{x}_2^k \cdot \boldsymbol{\theta}_2) \cdot f(\boldsymbol{x}_3^k \cdot \boldsymbol{\theta}_3),$$

[0567] 其中x_2^k、x_3^k是鉴别的HLA等位基因h＝2、h＝3的等位基因相互作用变量，并且θ_2、θ_3是测定的HLA等位基因h＝2、h＝3的参数的集合。

[0568] 又如，在使用网络变换函数$g_h(\bullet)$、$g_w(\bullet)$鉴别的m＝4种不同的HLA等位基因当中，肽p^k将由HLA等位基因h＝2、h＝3呈递的可能性可以由下式得到：

[0569] $$u_k = f(NN_2(\boldsymbol{x}_2^k;\boldsymbol{\theta}_2)) + f(NN_3(\boldsymbol{x}_3^k;\boldsymbol{\theta}_3)) - f(NN_2(\boldsymbol{x}_2^k;\boldsymbol{\theta}_2)) \cdot f(NN_3(\boldsymbol{x}_3^k;\boldsymbol{\theta}_3)),$$

[0570] 其中$NN_2(\bullet)$、$NN_3(\bullet)$是鉴别的HLA等位基因h＝2、h＝3的网络模型，并且θ_2、θ_3是测定的HLA等位基因h＝2、h＝3的参数的集合。

[0571] X.实施例5：预测模块

[0572] 预测模块320接收序列数据并使用呈递模型在序列数据中选择候选新抗原。确切地说，序列数据可以是从患者的肿瘤组织细胞中提取的DNA序列、RNA序列和/或蛋白质序列。预测模块320将序列数据处理成对于MHC-I具有8-15个氨基酸或对于MHC-II具有6-30个氨基酸的多个肽序列p^k。举例来说，预测模块320可以将给定序列“IEFROEIFJEF”处理成具有9个氨基酸的三个肽序列“IEFROEIFJ”、“EFROEIFJE”和“FROEIFJEF”。在一个实施方案中，预测模块320可以通过将从患者的正常组织细胞提取的序列数据与从患者的肿瘤组织细胞提取的序列数据相比较以鉴别含有一个或多个突变的部分，由此鉴别出作为突变肽序列的候选新抗原。

[0573] 预测模块320将一个或多个呈递模型应用于处理的肽序列以估计这些肽序列的呈递可能性。确切地说，预测模块320可以通过将呈递模型应用于候选新抗原来选择一个或多个可能被呈递于肿瘤HLA分子上的候选新抗原肽序列。在一个实施方式中，预测模块320选出估计呈递可能性超过预定阈值的候选新抗原序列。在另一个实施方式中，呈递模块选出v个具有最高估计呈递可能性的候选新抗原序列(其中v一般是可以在疫苗中递送的表位的最大数量)。包括选择用于给定患者的候选新抗原的疫苗可以注射到患者体内以诱导免疫反应。

[0574] XI.实施例6：盒设计模块

[0575] XI.A.概述

[0576] 盒设计模块324基于用于注射到患者体内的v所选候选物肽来生成疫苗盒序列。确切地说，对于包含在具有能力v的疫苗中的所选肽集合p^k，k＝1,2,…,v，通过使各自包含相应肽p^k的序列的一系列治疗性表位序列p'^k，k＝1,2,…,v串接来给予盒序列。盒设计模块324可使表位直接彼此相邻地串接。例如，疫苗盒C可被表示为：

[0577] $$\mathbf{C} = [\boldsymbol{p}'^{t_1}\ \boldsymbol{p}'^{t_2}\ \cdots\ \boldsymbol{p}'^{t_v}] \qquad (24)$$

[0578] 其中p'^{t_i}表示所述盒的第i个表位。因此，t_i对应于所选肽在盒的第i个位置处的指

数k＝1、2、…、v。盒设计模块324可通过相邻表位之间的一个或多个任选连接子使表位串接。例如，疫苗盒C可被表示为：

[0579] $$C=[p'^{t_1}\ l_{(t_1,t_2)}\ p'^{t_2}\ l_{(t_2,t_3)}\cdots\ l_{(t_{v-1},t_v)}\ p'^{t_v}] \quad (25)$$

[0580] 其中$l_{(ti,tj)}$表示置于盒的第i个表位p'^{ti}与第j＝i+1个表位$p'^{j=i+1}$之间的连接子序列。盒设计模块324确定所选表位p'^{k}，k＝1,2,…,v中的哪些布置在盒的不同位置以及置于表位之间的任何连接子序列。盒序列C可基于本说明书中所述的任何方法作为疫苗装载。

[0581] 所述治疗性表位集合可基于通过与高于预定阈值的呈递可能性相关联的预测模块320确定的所选肽来生成，其中呈递可能性通过呈递模型来确定。然而，应了解的是，在其他实施方案中，所述治疗性表位集合可基于许多方法中的任何一种或多种(单独或组合)，例如基于与患者的I类或II类HLA等位基因的结合亲和力或预测的结合亲和力、与患者的I类或II类HLA等位基因的结合稳定性或预测的结合稳定性、随机取样等来生成。

[0582] 在一个实施方案中，治疗性表位p'^{k}可对应于所选肽p^k本身。治疗性表位p'^{k}也可包含C末端和/或N末端侧接序列以及所选肽。例如，盒中所包含的表位p'^{k}可表示为序列[n^k p^k c^k]，其中c^k为连接所选肽p^k的C末端的C末端侧接序列，并且n^k是连接至所选肽p^k的N末端的N末端侧接序列。在本说明书其余部分通篇提到的一种情况下，N末端侧接序列和C末端侧接序列是治疗性免疫表位在其来源背景下的天然N末端侧接序列和C末端侧接序列。在本说明书其余部分通篇提到的一种情况下，治疗性表位p'^{k}表示固定长度的表位。在另一种情况下，治疗性表位p'^{k}可表示可变长度的表位，其中表位的长度可根据例如C或N侧接序列的长度而变化。例如，C末端侧接序列c^k和N末端侧接序列n^k各自可具有2-5个残基的可变长度，从而使得表位p'^{k}有16个可能的选择。

[0583] 盒设计模块324通过考虑跨越盒中一对治疗性表位之间的接合部的接合表位的呈递来生成盒序列。接合表位是新型非自身但不相关的表位序列，其由于使盒中的治疗性表位和连接子序列串接的过程而在盒中产生。接合表位的新型序列不同于盒本身的治疗性表位。跨越表位p'^{ti}和p'^{tj}的接合表位可包含与不同于治疗性表位p'^{ti}和p'^{tj}本身的序列的p'^{ti}或p'^{tj}二者重叠的任何表位序列。确切地说，盒的表位p'^{ti}与相邻表位p'^{tj}之间的具有或不具有任选连接子序列$l^{(ti,tj)}$的每个接合部可与$n_{(ti,tj)}$个接合表位$e_n{}^{(ti,tj)}$，n＝1,2,…,$n_{(ti,tj)}$缔合。接合表位可为至少部分与表位p'^{ti}和p'^{tj}二者重叠的序列，或者可为至少部分与置于表位p'^{ti}和p'^{tj}之间的连接子序列重叠的序列。接合表位可由I类MHC、II类MHC或二者呈递。

[0584] 图13示出两个示例性盒序列盒1(C_1)和盒2(C_2)。每个盒具有v＝2的疫苗能力，并且包含治疗性表位p'^{t1}＝p^1＝SINFEKL和p'^{t2}＝p^2＝LLLLLVVVV以及两个表位之间的连接子序列$1^{(t1,t2)}$＝AAY。确切地说，盒C_1的序列由[p^1 $1^{(t1,t2)}$ p^2]得到，而盒C_2的序列由[p^2 $1^{(t1,t2)}$ p^1]得到。盒C_1的示例性接合表位$e_n{}^{(1,2)}$可为跨越盒中的表位p'^{1}和p'^{2}二者的序列诸如EKLAAYLLL、KLAAYLLLLL和FEKLAAYL，并且可为跨越盒中的连接子序列和单一所选表位的序列诸如AAYLLLLL和YLLLLLVVV。类似地，盒C_2的示例性接合表位$e_m{}^{(2,1)}$可为诸如VVVVAAYSIN、VVVVAAY和AYSINFEK的序列。尽管两个盒涉及同一组序列p^1、$1^{(c1,c2)}$和p^2，所鉴别的接合表位集合根据盒内治疗性表位的排序序列而不同。

[0585] 盒设计模块324生成减小接合表位在患者中呈递的可能性的盒序列。确切地说，在将盒注入患者体内时，接合表位可能由患者的I类HLA或II类HLA等位基因呈递，并分别刺激

CD8或CD4 T细胞反应。此类反应经常是不希望的，因为与接合表位反应的T细胞不具有治疗益处，并且可通过抗原竞争减小对盒中选定的治疗性表位的免疫反应。[76]

[0586] 在一个实施方案中，盒设计模块324迭代遍历一个或多个候选盒，并且确定与盒序列缔合的接合表位的呈递评分低于数字阈值的该盒序列。接合表位呈递评分是与盒中的接合表位的呈递可能性相关联的数量，并且接合表位呈递评分的较高值指示盒的接合表位将由I类HLA或II类HLA或二者呈递的可能较高。

[0587] 盒设计模块324可确定候选盒序列中与最大接合表位呈递评分相关联的盒序列或者选择具有低于预定阈值的呈递评分的盒序列。在一种情况下，给定盒序列C的呈递评分是基于各自与盒C中的接合部相关联的距离度量集合$d(e_n^{(ti,tj)}, n=1,2,\cdots,n_{(ti,tj)}) = d_{(ti,tj)}$来确定的。确切地说，距离度量$d_{(ti,tj)}$指示将呈递跨越一对相邻治疗性表位$p'^{ti}$和$p'^{tj}$之间的一个或多个接合表位的可能性。盒C的接合表位呈递评分然后可通过将一个函数(例如，求和、统计函数)应用于盒C的距离度量集合来确定。在数学上，呈递评分由以下得到：

[0588] $$\text{评分} = h(d_{(t_1,t_2)}, d_{(t_2,t_3)}, \ldots, d_{(t_{v-1},t_v)}) \quad (26)$$

[0589] 其中h(•)为将每个接合部的距离度量映射至一个评分的一些函数。在本说明书其余部分通篇提到的一种具体情况下，函数h(•)是对盒的距离度量的求和。

[0590] 盒设计模块324可迭代遍历一个或多个候选盒序列，确定候选盒的接合表位呈递评分，并鉴别与低于阈值的接合表位呈递评分相关联的最佳盒序列。在本说明书其余部分通篇提到的一个具体实施方案中，给定接合部的距离度量d(•)可由呈递的接合表位的呈递可能性或预期数目的总和得到，如通过说明书的VII和VIII部分所述的呈递模型确定的。然而，将了解的是，在其他实施方案中，距离度量可由单独的其他因素或这些因素与如上文列举一个模型的模型组合得到，其中这些其他因素可包括由以下任何一种或多种(单独或组合)得到距离度量：I类HLA或II类HLA的HLA结合亲和力或稳定性测量或预测以及I类HLA或II类HLA的在HLA质谱法上训练的呈递或免疫原性模型或T细胞表位数据。例如，距离度量可将关于I类HLA和II类HLA呈递的信息组合。例如，距离度量可为被预测以低于阈值的结合亲和力结合患者I类HLA或II类HLA等位基因中的任一者的接合表位的数目。在另一个实例中，距离度量可为被预测由患者I类HLA或II类HLA等位基因中的任一者呈递的接合表位的预期数目。

[0591] 盒设计模块324可进一步检查一个或多个候选盒序列以确定候选盒序列中的任一接合表位是否是正设计疫苗所针对的给定患者的自身表位。为了完成此检查，盒设计模块324检查已知数据库诸如BLAST中的接合表位。在一个实施方案中，盒设计模块可被配置为通过将多对表位t_i, t_j的距离度量$d_{(ti,tj)}$设置成极大值(例如，100)来设计避免接合自身表位的盒，其中使表位t_i串接至表位t_j的N末端导致形成接合自身表位。

[0592] 返回至图13中的实施例，盒设计模块324确定(例如)盒C_1中的单一接合部(t_1, t_2)的距离度量$d_{(t1,t2)} = d_{(1,2)} = 0.39$，所述距离度量通过对长度为例如I类MHC的8至15个氨基酸或II类MHC的9-30个氨基酸的所有可能的接合表位$e_n^{(t1,t2)} = e_n^{(1,2)}$的呈递可能性求和来得到。由于盒$C_1$中不存在其他接合部，对盒$C_1$的距离度量求和的接合表位呈递评分也由0.39得到。盒设计模块324也确定盒C_2中的单一接合部的距离度量$d_{(t1,t2)} = d_{(2,1)} = 0.068$，所述距离度量通过对长度为I类MHC的8至15个氨基酸或II类MHC的9-30个氨基酸的所有可能的接合表位$e_n^{(t1,t2)} = e_n^{(2,1)}$的呈递可能性求和来得到。在此实施例中，盒$C_2$的接合表位呈递评

分也由单一接合部的距离度量0.068得到。盒设计模块324在接合表位呈递评分低于C_1的盒序列时输出作为最佳盒的C_2的盒序列。

[0593] 盒设计模块324可执行强力方法并迭代遍历所有活大部分可能的候选盒序列以选择具有最小接合表位呈递评分的序列。然而，此类候选盒的数目可随着疫苗v的能力增加而过大。例如，对于v＝20个表位的疫苗能力，盒设计模块324必须迭代遍历约10^{18}个可能的候选盒以确定具有最低接合表位呈递评分的盒。此确定对于盒设计模块324在合理时间量内完成以生成用于患者的疫苗而言可为计算繁重的(在需要的计算处理资源方面)且有时是可迭代的。此外，说明每个候选盒的可能接合表位可能甚至更繁重。因此，盒设计模块324可基于迭代遍历许多候选盒序列的方式来选择盒序列，所述候选盒序列的数目显著小于用于强力方法的候选盒序列的数目。

[0594] 在一个实施方案中，盒设计模块324生成随机或至少假随机生成的候选盒的子集，并且选择与低于预定阈值的接合表位呈递评分相关联的候选盒作为盒序列。另外，盒设计模块324可从具有最低接合表位呈递评分的子集中选择候选盒作为盒序列。例如，盒设计模块324可生成v＝20所选表位集合的约1百万个候选盒的子集，并且选择具有最小接合表位呈递评分的候选盒。尽管生成随机盒序列的子集并从所述子集中选择具有低接合表位呈递评分的盒序列相对于强力方法可能为次最佳的，但是这需要显著更少计算的资源，从而使其实施在技术上可行。此外，与这种更有效的技术相反，执行强力方法可仅产生微小或甚至可忽略的接合表位呈递评分的改进，因此从资源分配角度来看所述方法并无价值。

[0595] 在另一个实施方案中，盒设计模块324通过将盒的表位序列作为不对称旅行商问题(TSP)进行公式化来确定改进的盒配置。给定节点和每对节点之间的距离的列表，TSP确定与最短总距离相关联的节点的序列，以仅访问每个节点一次并返回至初始节点。例如，给定彼此之间的距离已知的城市A、B和C，TSP的解生成闭合的城市顺序，仅旅行访问每个城市一次的总距离是可能的路径中最小的。当一对节点之间的距离为不对称的时，TSP的不对称版本确定节点的最佳顺序。例如，从节点A旅行至节点B的“距离”可不同于从节点B旅行至节点A的“距离”。

[0596] 盒设计模块324通过解出不对称TSP来确定改进的盒序列，其中每个节点对应于治疗性表位p'^{k}。从对应于表位p'^{k}的节点到对应于表位p'^{m}的另一个节点的距离由接合表位距离度量$d_{(k,m)}$得到，而从对应于表位p'^{m}的节点到对应于表位p'^{k}的节点的距离由可不同于距离度量$d_{(k,m)}$的距离度量$d_{(m,k)}$得到。通过使用不对称TSP解出改进的最佳盒，盒设计模块324可找到导致穿过盒的表位之间的接合部的呈递评分减小的盒序列。不对称TSP的解指示治疗性表位的序列对应于表位应在盒中串接的排序，以使穿过盒的接合部的接合表位呈递评分最小化。确切地说，给定治疗性表位集合k＝1,2,…,v，盒设计模块324确定盒中每对可能排序的治疗性表位的距离度量$d_{(k,m)}$，k,m＝1,2,…,v。换句话说，对于给定的一对k,m表位，确定使治疗性表位p'^{m}串接在表位p'^{k}之后的距离度量$d_{(k,m)}$和使治疗性表位p'^{k}串接在表位p'^{m}之后的距离度量$d_{(m,k)}$，因为这些距离度量可彼此不同。

[0597] 盒设计模块324通过整数线性规划问题解出不对称TSP。确切地说，盒设计模块324生成由以下得到的(v+1)×(v+1)路径矩阵P：

[0598] $$\mathbf{P}=\begin{bmatrix}0 & \mathbf{0}^{1\times v}\\ \mathbf{0}^{v\times 1} & \mathbf{D}\end{bmatrix} \qquad (26)。$$

[0599] v x v矩阵D是不对称距离矩阵，其中每个元素D(k,m)，k＝1,2,…,v;m＝1,2,…,v对应于从表位p'^{k}到表位p'^{m}的接合部的距离度量。P的行k＝2,…,v对应于初始表位的节点，而第1行第1列对应于距离其他节点零距离的“幻象节点(ghost node)”。将“幻象节点”添加至矩阵编译了疫苗盒为线性而非环状的概念，因此在第一表位与最后表位之间不存在接合部。换句话说，序列不是环状的，并且第一表位并未假设串接在序列中的最后表位之后。使x_{km}表示二元变量，在表位p'^{k}串接至表位p'^{m}的N末端的情况下如果存在有向路径(即，盒中的表位-表位接合部)，则所述二元变量为1，并且反之则为0。另外，使E表示所有v个治疗性疫苗表位的集合，并且使$S \subset E$表示表位的子集。对于任何此类子集S，使out(S)表示表位-表位接合部x_{km}＝1的数目，其中k是S中的表位并且m是E\S中的表位。给定已知路径矩阵P，盒设计模块324找到解出以下整数线性规划问题的路径矩阵X:

[0600] $$\min_{x} \sum_{k=1}^{v+1} \sum_{k\neq m, m=1}^{v+1} P_{km} \cdot x_{km} \qquad (27)$$

[0601] 其中P_{km}表示经受以下限制的路径矩阵P的元素P(k,m)：

[0602] $$\sum_{k=1}^{v+1} x_{km} = 1, \quad m = 1,2,\ldots,v+1$$

[0603] $$\sum_{m=1}^{v+1} x_{km} = 1, \quad k = 1,2,\ldots,v+1$$

[0604] x_{kk}＝0,k＝1,2,…,v+1

[0605] $$\text{out}(S) \geq 1, \quad S \subset E, 2 \leq |S| \leq |V|/2$$

[0606] 前两个限制保证每个表位在盒中仅出现一次。最后限制确保盒连接。换句话说，由x编码的盒是连接的线性蛋白序列。

[0607] 等式(27)的整数线性规划问题中x_{km},k,m＝1,2,…,v+1的解指示闭合的节点和幻象节点顺序可用于推断盒的治疗性表位的一个或多个序列降低接合表位的呈递评分。确切地说，x_{km}＝1的值指示存在从节点k至节点m的“路径”，或者换句话说，在改进的盒序列中治疗性表位p'^{m}应串接在治疗性表位p'^{k}之后。x_{km}＝0的解指示不存在此类路径，或者换句话说，在改进的盒序列中治疗性表位p'^{m}不应串接在治疗性表位p'^{k}之后。总之，等式(27)的整数规划问题中x_{km}的值表示节点和幻象节点的顺序，其中输入路径并且仅一次存在每个节点。例如，$x_{幻象,1}$＝1、x_{13}＝1、x_{32}＝1和$x_{2,幻象}$＝1(反之则为0)的值可指示节点和幻象节点的顺序幻象→1→3→2→幻象。

[0608] 一旦已解出所述顺序，就从所述顺序中删除幻象节点以生成细化顺序，其中仅初始节点对应于盒中的治疗性表位。细化顺序指示所选表位应在盒中串接以改进呈递评分的排序。例如，从先前段落中的实施例继续，可删除幻象节点以生成细化顺序1→3→2。细化顺序指示串接盒中的表位的一种可能的方式，即p^{1}→p^{3}→p^{2}。

[0609] 当治疗性表位p'^{k}是可变长度的表位时，盒设计模块324确定对应于治疗性表位p'^{k}和p'^{m}的不同长度的候选距离度量，并且鉴别距离度量$d_{(k,m)}$为最小候选距离度量。例如，表位p'^{k}＝[n^{k} p^{k} c^{k}]和p'^{m}＝[n^{m} p^{m} c^{m}]可各自包含从(在一个实施方案中)2-5个氨基酸变化

的相应N末端侧接序列和C末端侧接序列。因此，表位p'^{k}与p'^{m}之间的接合部与16组不同的接合表位缔合，所述接合表位基于置于接合部中的n^{k}的4个可能长度值和c^{m}的4个可能长度值。盒设计模块324可确定每组接合表位的候选距离度量，并且确定距离度量$d_{(k,m)}$为最小值。盒设计模块324然后可构建路径矩阵P并解出等式(27)中的整数线性规划问题以确定盒序列。

[0610] 与随机取样方法相比，使用整数规划问题解出盒序列需要确定各自对应于疫苗中的一对治疗性表位的v x(v-1)距离度量。通过此方法确定的盒序列可导致与随机取样方法(尤其是所生成的候选盒序列的数目较大时)相比具有显著更少接合表位呈递，同时可能需要显著更少计算资源的序列。

[0611] XI.B.通过随机取样对比不对称TSP生成的盒序列的接合表位呈递的比较

[0612] 通过随机取样1,000,000个置换(盒序列C_1)并通过解出等式(27)中的整数线性规划问题(盒序列C_2)来生成包含v＝20个治疗性表位的两个盒序列。距离度量以及因此的呈递评分基于等式(14)中所述的呈递模型来确定，其中f是s型函数，x_h^i是太p^i的序列，$g_h(\bullet)$是神经网络函数，w包含侧接序列、log肽p^i的转录物/百万千碱基(TPM)、肽p^i的蛋白的免疫原性和肽p^i的来源的样品ID，并且侧接序列和log TPM的$g_w(\bullet)$分别是神经网络函数。$g_h(\bullet)$的每个神经网络函数包括一个隐藏层的多层感知器(MLP)的一个输出节点，所述多层感知器具有输出尺寸231(11个残基x21个字符/残基，包括填充字符)、宽度256、隐藏层中的修正线性单元(ReLU)活化、输出层中的线性活化和训练数据集中每个HLA等位基因的一个输出节点。侧接序列的神经网络函数是一个隐藏层的MLP，其具有输出尺寸210(N末端侧接序列的5个残基+C末端侧接序列的5个残基x21个字符/残基，包括填充字符)、宽度32、隐藏层中的ReLU活化和输出层中的线性活化。RNA log TPM的神经网络函数是一个隐藏层的MLP，其具有输出尺寸1、宽度16、隐藏层中的ReLU活化和输出层中的线性活化。构建HLA等位基因HLA-A*02:04、HLA-A*02:07、HLA-B*40:01、HLA-B*40:02、HLA-C*16:02和HLA-C*16:04的呈递模型。比较指示两个盒序列的呈递的接合表位的预期数目的呈递评分。结果显示通过解出等式(27)生成的盒序列的呈递评分与相对于通过随机取样生成的盒序列的呈递评分的约4倍改进相关联。

[0613] 确切地说，v＝20个表位由以下得出：

[0614] p'^{1}＝YNYSYWISIFAHTMWYNIWHVQWNK

[0615] p'^{2}＝IEALPYVFLQDQFELRLLKGEQGNN

[0616] p'^{3}＝DSEETNTNYLHYCHFHWTWAQQTTV

[0617] p'^{4}＝GMLSQYELKDCSLGFSWNDPAKYLR

[0618] p'^{5}＝VRIDKFLMYVWYSAPFSAYPLYQDA

[0619] p'^{6}＝CVHIYNNYPRMLGIPFSVMVSGFAM

[0620] p'^{7}＝FTFKGNIWIEMAGQFERTWNYPLSL

[0621] p'^{8}＝ANDDTPDFRKCYIEDHSFRFSQTMN

[0622] p'^{9}＝AAQYIACMVNRQMTIVYHLTRWGMK

[0623] p'^{10}＝KYLKEFTQLLTFVDCYMWITFCGPD

[0624] p'^{11}＝AMHYRTDIHGYWIEYRQVDNQMWNT

[0625] p'^{12}＝THVNEHQLEAVYRFHQVHCRFPYEN

[0626] p'^{13}＝QTFSECLFFHCLKVWNNVKYAKSLK

[0627] p'^{14}＝SFSSWHYKESHIALLMSPKKNHNNT

[0628] p'^{15}＝ILDGIMSRWEKVCTRQTRYSYCQCA

[0629] p'^{16}＝YRAAQMSKWPNKYFDFPEFMAYMPI

[0630] p'^{17}＝PRPGMPCQHHNTHGLNDRQAFDDFV

[0631] p'^{18}＝HNIISDETEVWEQAPHITWVYMWCR

[0632] p'^{19}＝AYSWPVVPMKWIPYRALCANHPPGT

[0633] p'^{20}＝HVMPHVAMNICNWYEFLYRISHIGR。

[0634] 在第一实施例中,随机生成具有20个治疗性表位的1,000,000个不同候选盒序列。生成每个候选盒序列的呈递评分。被鉴别为具有最低呈递评分的候选盒序列未:

[0635] C_1＝THVNEHQLEAVYRFHQVHCRFPYENAMHYQMWNTYRAAQMSKWPNKYFDFPEFMAYMPICVHIY NNYPRMLGIPFSVMVSGFAMAYSWPVVPMKWIPYRALCANHPPGTANDDTPDFRKCYIEDHSFRFSQTMNIEALPYV FLQDQFELRLLKGEQGNNDSEETNTNYLHYCHFHWTWAQQTTVILDGIMSRWEKVCTRQTRYSYCQCAFTFKGNIWI EMAGQFERTWNYPLSLSFSSWHYKESHIALLMSPKKNHNNTQTFSECLFFHCLKVWNNVKYAKSLKHVMPHVAMNIC NWYEFLYRISHIGRHNIISDETEVWEQAPHITWVYMWCRVRIDKFLMYVWYSAPFSAYPLYQDAKYLKEFTQLLTFV DCYMWITFCGPDAAQYIACMVNRQMTIVYHLTRWGMKYNYSYWISIFAHTMWYNIWHVQWNKGMLSQYELKDCSLGF SWNDPAKYLRPRPGMPCQHHNTHGLNDRQAFDDFV

[0636] 其中呈递评分6.1预期呈递的接合表位的数目。1,000,000个随机序列的中值呈递评分为18.3。实验显示呈递的接合表位的预期数目可通过鉴别随机取样的盒内的盒序列来显著减少。

[0637] 在第二实施例中,盒序列C_2通过解出等式(27)中的整数线性规划问题来鉴别。确切地说,确定一对治疗性表位之间的每个潜在接合部的距离度量。使用距离度量解出整数规划问题的解。通过此方法鉴别的盒序列为:

[0638] C_2＝IEALPYVFLQDQFELRLLKGEQGNNILDGIMSRWEKVCTRQTRYSYCQCAHVMPHVAMNICNWY EFLYRISHIGRTHVNEHQLEAVYRFHQVHCRFPYENFTFKGNIWIEMAGQFERTWNYPLSLAMHYQMWNTSFSSWHY KESHIALLMSPKKNHNNTVRIDKFLMYVWYSAPFSAYPLYQDAQTFSECLFFHCLKVWNNVKYAKSLKYRAAQMSKW PNKYFDFPEFMAYMPIAYSWPVVPMKWIPYRALCANHPPGTCVHIYNNYPRMLGIPFSVMVSGFAMHNIISDETEVW EQAPHITWVYMWCRAAQYIACMVNRQMTIVYHLTRWGMKYNYSYWISIFAHTMWYNIWHVQWNKGMLSQYELKDCSL GFSWNDPAKYLRKYLKEFTQLLTFVDCYMWITFCGPDANDDTPDFRKCYIEDHSFRFSQTMNDSEETNTNYLHYCHF HWTWAQQTTVPRPGMPCQHHNTHGLNDRQAFDDFV

[0639] 其中呈递评分为1.7。盒序列C_2的呈递评分显示相对于盒序列C_1的呈递评分的约4倍改进和相对于1,000,000个随机生成的候选盒的中值呈递评分的约11倍改进。生成盒C1的运行时间在2.30GHz Intel Xeon E5-2650 CPU的单线螺纹上是20秒。生成盒C_2的运行时间在相同CPU的单线螺纹上是1秒。因此,在此实施例中,通过解出等式(27)的整数规划问题鉴别的盒序列在20倍减少的计算成本下产生好约4倍的解。

[0640] 结果显示整数规划问题可潜在提供呈递的接合表位数目低于由随机取样鉴别的数目、潜在具有更少计算资源的盒序列。

[0641] XI.C.通过MHCflurry和呈递模型生成的盒序列选择的接合表位呈递的比较

[0642] 在此实施例中,基于肿瘤/正常外显子测序、肿瘤转录组测序和肺癌样品的HLA分型选择包含v＝20个治疗性表位的盒序列,通过随机取样1,000,000个置换并通过解出等式

(27)中的整数线性规划问题来生成所述盒序列。距离度量以及因此的呈递评分基于由HLA-肽结合亲和力预测器MHCflurry预测的接合表位的数目来确定，以便以低于多个阈值(例如，50-1000nM或更高或更低)的亲和力结合患者的HLA。在此实施例中，根据以上XI.B部分中的呈递模型通过对突变进行排名从肿瘤样品中鉴别的98个体细胞突变选择20种被选择为治疗性表位的非同义体细胞突变。然而，应了解的是，在其他实施方案中，可基于其他标准，诸如基于稳定性的标准或诸如呈递评分、亲和力等标准的组合来选择治疗性表位。另外，应了解的是，用于对包含在疫苗中的治疗性表位进行优先级排序的标准不需要与用于确定盒设计模块324中所用的距离度量D(k,m)的标准相同。

[0643] 患者的I类HLA等位基因为HLA-A*01:01、HLA-A*03:01、HLA-B*07:02、HLA-B*35:03、HLA-C*07:02、HLA-C*14:02。

[0644] 确切地说，在此实施例中，v＝20个治疗性表位为

[0645]
SSTPYLYYGTSSVSYQFPMVPGGDR
EMAGKIDLLRDSYIFQLFWREAAEP
ALKQRTWQALAHKYNSQPSVSLRDF
VSSHSSQATKDSAVGLKYSASTPVR
KEAIDAWAPYLPEYIDHVISPGVTS
SPVITAPPSSPVFDTSDIRKEPMNI
PAEVAEQYSEKLVYMPHTFFIGDHA
MADLDKLNIHSIIQRLLEVRGS
AAAYNEKSGRITLLSLLFQKVFAQI
KIEEVRDAMENEIRTQLRRQAAAHT
DRGHYVLCDFGSTTNKFQNPQTEGV
QVDNRKAEAEEAIKRLSYISQKVSD
CLSDAGVRKMTAAVRVMKRGLENLT
LPPRSLPSDPFSQVPASPQSQSSSQ
ELVLEDLQDGDVKMGGSFRGAFSNS
VTMDGVREEDLASFSLRKRWESEPH
IVGVMFFERAFDEGADAIYDHINEG
TVTPTPTPTGTQSPTPTPITTTTTV
QEEMPPRPCGGHTSSSLPKSHLEPS
PNIQAVLLPKKTDSHHKAKGK

[0646] 下表中来自此实施例的结果比较了由MHCflurry预测的以低于阈值列(其中nM表示纳摩尔)中的值的亲和力结合患者的HLA的接合表位的数目，如经由三种示例性方法发现的。对于第一种方法，经由上文所述的旅行商问题(ATSP)公式以1s运行时间发现的最佳盒。对于第二种方法，如通过利用在1百万随机样品之后发现的最好盒确定的最佳盒。对于第三种方法，在1百万随机样品中发现中值数目的接合表位。

[0647]

阈值 (nM)	ATSP #结合接合表位	随机取样 #结合接合表位	中值 #结合接合表位
50	0	0	3
100	0	0	7
150	0	1	12
500	15	26	55
1000	68	91	131

[0648] 此实施例的结果说明许多标准中的任一种可用于鉴别给定盒设计是否满足设计要求。确切地说，如通过先前实施例证实的，从许多候选物中选出的盒序列可通过具有最低

接合表位呈递评分或至少低于鉴别的阈值的评分的盒序列来指定。此实施例表明，另一种标准诸如结合亲和力可用于鉴别给定盒设计是否满足设计要求。对于此标准，阈值结合亲和力(例如50-1000或更大或更低)可为指示盒设计序列应具有少于一些阈值数目的高于阈值(例如，0)的接合表位的集合，并且可使用许多方法中的任一种(例如，表中所示出的一至三种方法)，其可用于鉴别给定候选盒序列是否满足那些要求。这些示例性方法进一步说明根据所用方法，阈值可能需要不同地设置。可以设想其他标准，诸如基于稳定性的那些标准或诸如呈递评分、亲和力等标准的组合。

[0649] 在另一个实施例中，使用此部分(XI.C)前面的相同HLA类型和20种治疗性表位生成相同盒，而不是使用基于结合亲和力预测的距离度量，表位m、k的距离度量是跨越被预测为有患者I类HLA等位基因呈递的m至k接合部的肽的数目，其呈递概率高于一系列阈值(0.005与0.5的概率之间或更高或更低)，其中呈递的概率他用过以上XI.B部分中的呈递模型来确定。此实施例进一步说明了在鉴别给定盒序列是否满足在疫苗中使用的设计要求时考虑的标准的宽度。

[0650]

阈值 (概率)	ATSP # 接合表位	随机取样 #接合表位	中值 #接合表位
0.005	58	79	118
0.01	39	59	93
0.05	7	33	47
0.1	5	14	35
0.2	1	8	25
0.5	0	2	14

[0651] 以上实施例已鉴别用于确定候选盒序列是否可通过实施来改变的标准。这些实施例各自已说明高于或低于所述标准的接合表位的数目的计数可为用于确定候选盒序列是否满足该标准的计数。例如，如果标准是满足或超过对HLA的阈值结合亲和力的表位的数目，则候选盒序列的数目是否高于该数目可确定候选盒序列是否满足用作疫苗的所选盒的标准。如果标准是超过阈值呈递可能性的接合表位的数目，则类似。

[0652] 然而，在其他实施方案中，可执行除了计数以外的计算以确定候选盒序列满足设计标准。例如，不计数超过/低于一些阈值的表位，反而确定接合表位超过或低于所述阈值的比例，例如接合表位的顶部X％具有高于一些阈值Y的呈递可能性，或者确定接合表位的X％百分比是否具有小于或大于Z nM的HLA结合亲和力。这些仅是实施例，通常标准可基于任一个别接合表位的任何属性或来源于一些或所有接合表位的聚集物的统计学。在此，X通常可为0与100％之间的任何数值(例如，75％或更小)并且Y可为0与1之间的任何值，并且Z可为适于所讨论的标准的任何数值。这些值可凭经验确定，并且依赖于所用的模型和标准以及所用训练数据的品质。

[0653] 这样，在某些方面，可去除具有高呈递概率的接合表位；可保留具有低呈递概率的接合表位；可去除紧密结合的接合表位，即具有低于1000nM或500nM或一些其他阈值的结合亲和力的接合表位；且/或可保留弱结合的接合表位，即具有高于1000nM或500nM或一些其他阈值的结合亲和力的接合表位。

[0654] 尽管以上实施例已使用上文所述呈递模型的实施方式鉴别候选序列，但是这些理论同样适用于盒序列中布置的表位也基于其他类型的模型(诸如基于亲和力、稳定性等)鉴

别的实施方式。

[0655] XI.D.共有抗原和共有新抗原的盒选择

[0656] 不选择用于个别患者的个人化疫苗的治疗性表位子集，一系列治疗性表位序列p'^{k}，k＝1,2,…,v可为与癌症患者群体中的高呈递可能性相关联的表位集合。例如，所述一系列治疗性表位序列可为共有抗原序列，其为来自被鉴别为在癌症患者中过表达的基因的序列，并且与癌症患者群体中的高呈递可能性相关联。又如，所述一系列治疗性表位序列可为共有新抗原序列，其为与癌症患者群体中的普通驱动物突变相关联的序列，并且与高呈递可能性相关联。因此，并未基于个别患者的测序数据和HLA等位基因类型来定制盒的治疗性表位序列，治疗性表位序列可为多个患者中共有的。

[0657] 当共有盒序列时，一对表位t_i和t_j之间的距离度量$d_{(ti,tj)}$可被确定为各自与相应HLA等位基因相关联的次距离度量的加权和。确切地说，距离度量$d_{(ti,tj)}$可由以下得到：

[0658] $$d_{(ti,tj)} = \sum_{h=1}^{m} w_h \cdot d'_{h,(ti,tj)} \qquad (28)$$

[0659] 其中$d_{h,(ti,tj)}$为指示跨越一对相邻治疗性表位之间的一个或多个接合表位$e_n^{(ti,tj)}$，n＝1,2,…,$n_{(ti,tj)}$将在HLA等位基因h上呈递的可能性的次距离度量，并且w_h是指示HLA等位基因h在给定患者群体中的流行性的加权。通过如等式(28)或以使用HLA等位基因的流行性加权接合表位呈递的任何其他类似方式设置距离度量，可选择减少对于据估计在患者群体中更流行的HLA等位基因的接合表位呈递的盒序列。

[0660] 与HLA等位基因h相关联的次距离度量可由呈递的接合表位在HLA等位基因h上的呈递可能性或预期数目的总和得到，如通过说明书的VII和VIII部分所述的呈递模型确定的。然而，将了解的是，在其他实施方案中，次距离度量可由单独的其他因素或这些因素与如上文列举一个模型的模型组合得到，其中这些其他因素可包括由以下任何一种或多种(单独或组合)得到次距离度量：I类HLA或II类HLA的HLA结合亲和力或稳定性测量或预测以及I类HLA或II类HLA的在HLA质谱法上训练的呈递或免疫原性模型或T细胞表位数据。次距离度量可将关于I类HLA和II类HLA呈递的信息组合。例如，次距离度量可为被预测以低于阈值的结合亲和力结合患者I类HLA或II类HLA等位基因中的任一者的接合表位的数目。在另一个实例中，次距离度量可为被预测由患者I类HLA或II类HLA等位基因中的任一者呈递的接合表位的预期数目。

[0661] 基于等式(28)中定义的距离度量，盒设计模块324可使用以上XI.A部分介绍的任何方法迭代遍历一个或多个候选盒序列，确定候选盒的接合表位呈递评分，并鉴别与低于阈值的接合表位呈递评分相关联的最佳盒序列。

[0662] XI.E.共有抗原和共有新抗原的通过随机取样对比不对称TSP生成的盒序列的接合表位呈递的比较

[0663] 在此实施例中，使用来自XI.C部分的相同20种治疗性表位生成盒，并且比较通过三种示例性方法发现的盒序列的接合表位的预期数目。不同于XI.C部分，使用等式(28)确定距离度量和距离矩阵。使用来自XI.B部分的模型训练样品穿过28个HLA-A、43个HLA-B和23个HLA-C等位基因计算等位基因频率，其在等式(28)中表示为w_h。这些等位基因是由模型支持的等位基因。单独计算每种基因HLA-A、HLA-B和HLA-C的频率。基于高于由相应等位基因频率在不同阈值概率下加权的阈值呈递可能性的呈递的接合表位的预期数目来确定每

个距离度量。与XI.B部分类似，对于第一种方法，经由上文所述的旅行商问题(ATSP)公式发现最佳盒。对于第二种方法，通过利用在1百万随机样品之后发现的最好盒确定最佳盒。对于第三种方法，在1百万随机样品中发现中值数目的接合表位。确切地说，ATSP方法的距离矩阵是由等位基因频率加权的单等位基因距离次矩阵的加权和。

[0664]

阈值(概率)	ATSP接合表位的预期#	随机取样接合表位的预期#	中值接合表位的预期#
0.005	64.4	82.7	112.3
0.01	46.2	62.2	86.2
0.05	18.0	25.3	41.5
0.1	10.0	16.5	27.5
0.2	5.4	8.8	16.7
0.5	1.4	3.0	6.5

[0665] 如上表所示，由于每种方法中的距离度量是基于等位基因频率的接合表位的加权预期，结果不再是如XI.C中整数取值，因为距离矩阵不再是整数取值。结果显示整数规划问题也可提供共有抗原或共有新抗原的盒序列，其与由随机取样鉴别的盒序列相比减少呈递的接合表位用于共有(新)抗原疫苗盒包装的机会，且潜在地具有更少计算资源。

[0666] 在另一个实施例中，使用来自XI.C部分的相同20种治疗性表位生成盒，并且使用MHCflurry比较通过三种示例性方法发现的盒序列的接合表位的预期数目。使用等式(28)确定距离度量和距离矩阵。使用模型训练样品穿过22个HLA-A、27个HLA-B和9个HLA-C等位基因计算等位基因频率，其在等式(28)中表示为w_h。单独计算每种基因HLA-A、HLA-B和HLA-C的频率。基于低于由相应等位基因频率在不同阈值概率下加权的阈值结合亲和力的呈递的接合表位的预期数目来确定每个距离度量。与XI.B部分类似，对于第一种方法，经由上文所述的旅行商问题(ATSP)公式发现最佳盒。对于第二种方法，通过利用在1百万随机样品之后发现的最好盒确定最佳盒。对于第三种方法，在1百万随机样品中发现中值数目的接合表位。确切地说，ATSP方法的距离矩阵是由等位基因频率加权的单等位基因距离次矩阵的加权和。

[0667]

阈值(nM)	ATSP结合接合表位的预期#	随机取样结合接合表位的预期#	中值结合接合表位的预期#
50	0.3	0.7	2.7
100	0.9	1.7	4.9
150	1.6	3.1	6.8
500	6.7	9.5	15.9
1000	12.9	17.4	26.1

[0668] 此实施例的结果说明许多标准中的任一种可用于鉴别给定盒设计是否满足设计要求。确切地说，此实施例表明，另一种标准诸如结合亲和力可用于指示给定盒设计是否满足共有抗原和新抗原疫苗盒的设计要求。对于此标准，阈值结合亲和力(例如50-1000或更大或更低)可为指示盒设计序列应具有少于一些阈值数目的高于阈值(例如，0)的接合表位的集合，并且可使用许多方法中的任一种(例如，表中所示出的一至三种方法)，其可用于鉴别给定候选盒序列是否满足那些要求。这些示例性方法进一步说明根据所用方法，阈值可能需要不同地设置。可以设想其他标准，诸如基于稳定性的那些标准或诸如呈递评分、亲和

力等标准的组合。

[0669] XII.示例性计算机

[0670] 图14示出用于实施图1和3中所示的实体的示例性计算机1400。计算机1400包括耦合至芯片组1404的至少一个处理器1402。芯片组1404包括内存控制器集线器1420和输入/输出(I/O)控制器集线器1422。内存1406和图形适配器1412耦合至内存控制器集线器1420，并且显示器1418耦合至图形适配器1412。存储装置1408、输入装置1414和网络适配器1416耦合至I/O控制器集线器1422。计算机1400的其他实施方案具有不同的架构。

[0671] 存储装置1408是非暂时性计算机可读存储介质，如硬盘驱动器、致密光盘只读存储器(CD-ROM)、DVD或固态内存装置。内存1406保存处理器1402所使用的指令和数据。输入接口1414是触摸屏界面、鼠标、轨迹球或其他类型的指向装置、键盘或其某一组合，并且用于将数据输入计算机1400中。在一些实施方案中，计算机1400可以被配置成通过用户的示意动作从输入接口1414接收输入(例如，命令)。图形适配器1412将图像和其他信息显示于显示器1418上。网络适配器1416将计算机1400耦合至一个或多个计算机网络。

[0672] 计算机1400被调适成执行计算机程序模块以提供本文所述的功能。如本文所使用，术语“模块”是指用于提供指定功能的计算机程序逻辑。因此，模块可以在硬件、固件和/或软件中实施。在一个实施方案中，程序模块被存储于存储装置1408上，装载至内存1406中并由处理器1402执行。

[0673] 图1的实体所使用的计算机1400的类型可以根据实施方案和实体所需的处理能力而变化。举例来说，呈递鉴别系统160可以在单一计算机1400或在通过网络，如在服务器群中彼此通信的多台计算机1400中运行。计算机1400可以缺少以上描述的组件中的一些，如图形适配器1412和显示器1418。

[0674] 参考文献

[0675] 1.Desrichard,A.,Snyder,A.&Chan,T.A.Cancer Neoantigens and Applications for Immunotherapy.Clin.Cancer Res.Off.J.Am.Assoc.Cancer Res.(2015).doi:10.1158/1078-0432.CCR-14-3175

[0676] 2.Schumacher,T.N.&Schreiber,R.D.Neoantigens in cancer immunotherapy.Science 348,69-74(2015).

[0677] 3.Gubin,M.M.,Artyomov,M.N.,Mardis,E.R.&Schreiber,R.D.Tumor neoantigens:building a framework for personalized cancer immunotherapy.J.Clin.Invest.125,3413-3421(2015).

[0678] 4.Rizvi,N.A.et al.Cancer immunology.Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer.Science 348,124-128(2015).

[0679] 5.Snyder,A.et al.Genetic basis for clinical response to CTLA-4 blockade in melanoma.N.Engl.J.Med.371,2189-2199(2014).

[0680] 6.Carreno,B.M.et al.Cancer immunotherapy.A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells.Science 348,803-808(2015).

[0681] 7.Tran,E.et al.Cancer immunotherapy based on mutation-specific CD4+ T

cells in a patient with epithelial cancer.Science 344,641-645(2014).

[0682] 8.Hacohen,N.&Wu,C.J.-Y.United States Patent Application:0110293637-COMPOSITIONS AND METHODS OF IDENTIFYING TUMOR SPECIFIC NEOANTIGENS.(A1).at<http://appft1.uspto.gov/netacgi/nph-Parser?Sect1＝PTO1&Sect2＝HITOFF&d＝PG01&p＝1&u＝/netahtml/PTO/srchnum.html&r＝1&f＝G&1＝50&s1＝20110293637.PGNR.>

[0683] 9.Lundegaard,C.,Hoof,I.,Lund,O.&Nielsen,M.State of the art and challenges in sequence based T-cell epitope prediction.Immunome Res.6 Suppl 2,S3(2010).

[0684] 10.Yadav,M.et al.Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing.Nature 515,572-576(2014).

[0685] 11.Bassani-Sternberg,M.,Pletscher-Frankild,S.,Jensen,L.J.&Mann,M.Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation.Mol.Cell.Proteomics MCP 14,658-673(2015).

[0686] 12.Van Allen,E.M.et al.Genomic correlates of response to CTLA-4 blockade in metastatic melanoma.Science 350,207-211(2015).

[0687] 13.Yoshida,K&Ogawa,S.Splicing factor mutations and cancer.Wiley Interdiscip.Rev.RNA 5,445-459(2014).

[0688] 14.Cancer Genome Atlas Research Network.Comprehensive molecular profiling of lung adenocarcinoma.Nature 511,543-550(2014).

[0689] 15Rajasagi,M.et al.Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia.Blood 124,453-462(2014).

[0690] 16.Downing,S.R.et al.United States Patent Application:0120208706-OPTIMIZATION OF MULTIGENE ANALYSIS OF TUMOR SAMPLES.(A1)at<http://appft1.uspto.gov/netacgi/nph-Parser?Sect1＝PTO1&Sect2＝HITOFF&d＝PG01&p＝1&u＝/netahtml/PTO/srchnum.html&r＝1&f＝G&1＝50&s1＝20120208706.PGNR.>

[0691] 17.Target Capture for NextGen Sequencing-IDT.at<http://www.idtdna.com/pages/products/nextgen/target-capture>

[0692] 18.Shukla,S.A.et al.Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes.Nat.Biotechnol.33,1152-1158(2015).

[0693] 19.Cieslik,M.et alThe use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing.Genome Res.25,1372-1381(2015).

[0694] 20.Bodini,M.et alThe hidden genomic landscape of acute myeloid leukemia:subclonal structure revealed by undetected mutations.Blood 125,600-605(2015).

[0695] 21.Saunders,C.T.et al.Strelka:accurate somatic small-variant calling from sequenced tumor-normal sample pairs.Bioinforma.Oxf.Engl.28,1811-1817(2012).

[0696] 22.Cibulskis,K.et alSensitive detection of somatic point mutations in impure and heterogeneous cancer samples.Nat.Biotechnol.31,213-219(2013).

[0697] 23.Wilkerson,M.D.et al.Integrated RNA and DNA sequencing improves mutation detection in low purity tumors.Nucleic Acids Res.42,e107(2014).

[0698] 24.Mose,L.E.,Wilkerson,M.D.,Hayes,D.N.,Perou,C.M.&Parker,J.S.ABRA:improved coding indel detection via assembly-based realignment.Bioinforma.Oxf.Engl.30,2813-2815(2014)

[0699] 25.Ye,K.,Schulz,M.H.,Long,Q.,Apweiler,R.&Ning,Z.Pindel:a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads.Bioinforma.Oxf.Engl.25,2865-2871(2009).

[0700] 26.Lam,H.Y.K.et al.Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library.Nat.Biotechnol.28,47-55(2010).

[0701] 27.Frampton,G.M.et al.Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing.Nat.Biotechnol.31,1023-1031(2013).

[0702] 28.Boegel,S.et al.HLA typing from RNA-Seq sequence reads.Genome Med.4,102(2012).

[0703] 29.Liu,C.et al.ATHLATES:accurate typing of human leukocyte antigen through exome sequencing.Nucleic Acids Res.41,e142(2013).

[0704] 30.Mayor,N.P.et al.HLA Typing for the Next Generation.PloS One 10,e0127153(2015).

[0705] 31.Roy,C.K.,Olson,S.,Graveley,B.R.,Zamore,P.D.&Moore,M.J.Assessing long-distance RNA sequence connectivity via RNA-templated DNA-DNA ligation.eLife 4,(2015).

[0706] 32.Song,L.&Florea,L.CLASS:constrained transcript assembly of RNA-seq reads.BMC Bioinformatics 14 Suppl 5,S14(2013).

[0707] 33.Maretty,L.,Sibbesen,J.A.&Krogh,A.Bayesian transcriptome assembly.Genome Biol.15,501(2014).

[0708] 34.Pertea,M.et al.String Tie enables improved reconstruction of a transcriptome from RNA-seq reads.Nat.Biotechnol.33,290-295(2015).

[0709] 35.Roberts,A.,Pimentel,H.,Trapnell,C.&Pachter,L.Identification of novel transcripts in annotated genomes using RNA-SeqBioinforma.Oxf.Engl.(2011).doi:101093/bioinformatics/btr355

[0710] 36.Vitting-Seerup,K.,Porse,B.T.,Sandelin,A.&Waage,J.spliceR:an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data.BMC Bioinformatics 15,81(2014).

[0711] 37.Rivas,M.Aet al.Human genomics.Effect of predicted protein-

truneating genetic variants on the human transcriptome.Science 348,666-669 (2015).

[0712] 38.Skelly,D.A.,Johansson,M.,Madeoy,J.,Wakefield,J.&Akey,J.M.A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data.Genome Res.21,1728-1737(2011).

[0713] 39.Anders,S.,Pyl,P.T.&Huber,W.HTSeq--aPython framework to work with high-throughput sequencing data.Bioinforma.Oxf.Engl.31,166-169(2015).

[0714] 40.Furney,S.J.et al.SF3B1 mutations are associated with alternative splicing in uveal melanoma.Cancer Discov.(2013).doi:10.1158/2159-8290.CD-13-0330

[0715] 41.Zhou,Q.et al.A chemical genetics approach for the fuctional assessment of novel cancer genes.Cancer Res.(2015).doi:10.1158/0008-5472.CAN-14-2930

[0716] 42.Maguire,S.L.et al.SF3B1 mutations constitute a novel therapeutic target in breast cancer.J.Pathol.235,571-580(2015).

[0717] 43.Carithers,L.J.el al.A Novel Approach to High-Quality Postmortem Tissue Procurement:The GTEx Project.Biopreservation Biobanking 13,311-319 (2015).

[0718] 44.Xu,G.et al.RNA CoMPASS:a dual approach for pathogen and host transcriptome analysis of RNA-seq datasets.PloS One 9,e89445(2014).

[0719] 45.Andreatta,M.&Nielsen,M.Gapped sequence alignment using artificial neural networks:application to the MHC class I system.Bioinforma.Oxf.Engl. (2015).doi:10.1093/bioinformatics/btv639

[0720] 46.Jørgensen,K.W.,Rasmussen,M.,Buus,S.&Nielsen,M.NetMHCstab-predicting stability of peptide-MHC-I complexes;impacts for cytotoxic T lymphocyte epitope discovery.Immunology 141,18-26(2014).

[0721] 47.Larsen,M.V.et al.An integrative approach to CTL epitope prediction:a combined algorithm integrating MHC class I binding,TAP transport efficiencv,and proteasomal cleavage predictions.Eur.J.Immunol.35,2295-2303 (2005).

[0722] 48.Nielsen,M.,Lundegaard,C.,Lund,O.&Keşmir,C.The role of the proteasome in generating cytotoxic T-cell epitopes:insights obtained from improved predictions of proteasomal cleavage.Immunogenetics 57,33-41(2005).

[0723] 49.Boisvert,F.-M.et al.A Quantitative Spatial Proteomics Analysis of Proteome Turnover in Human Cells.Mol.Cell.Proteomics 11,M111.011429-M111.011429(2012).

[0724] 50Duan,F.et al.Genomic and bioinfornatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity.J.Exp.Med.211,2231-2248(2014).

[0725] 51.Janeway's Immunobiology:9780815345312:Medicine&Health Science Books@Amazon.com.at<http://www.amazon.com/Janeways-Immunobiology-Kenneth-Murphy/dp/0815345313>

[0726] 52.Calis,J.J.A.et al.Properties of MHC Class I Presented Peptides That Enhance Immunogenicity.PLoS Comput.Biol.9,e1003266(2013).

[0727] 53.Zhang,J.et al.Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing.Science 346,256-259 (2014)

[0728] 54.Walter,M.J.et al.Clonal architecture of secondary acute myeloid leukemia.N.Engl.J.Med.366,1090-1098(2012).

[0729] 55.Hunt DF,Henderson RA,Shabanowitz J,Sakaguchi K,Michel H,Sevilir N,Cox AL,Appella E,Engelhard VH.Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry.Science 1992.255:1261-1263.

[0730] 56.Zarling AL,Polefrone JM,Evans AM,Mikesh LM,Shabanowitz J,Lewis ST,Engelhard VH,Hunt DF.Identification of class I MHC-associated phosphopeptides as targets for cancer immunotherapy.Proc Natl Acad Sci USA.2006 Oct 3;103 (40):14889-94.

[0731] 57.Bassani-Sternberg M,Pletscher-Frankild S,Jensen LJ,Mann M.Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation.Mol Cell Proteomics.2015 Mar;14(3):658-73.doi:10.1074/mcp.M114.042812.

[0732] 58.Abelin JG,Trantham PD,Penny SA,Patterson AM,Ward ST,Hildebrand WH,Cobbold M,Bai DL,Shabanowitz J,Hunt DF.Complementary IMAC enrichment methods for HLA-associated phosphopeptide identification by mass spectrometry.Nat Protoc.2015 Sep;10(9):1308-18.doi:10.1038/nprot.2015.086.Epub 2015 Aug 6

[0733] 59.Barnstable CJ,Bodmer WF,Brown G,Galfre G,Milstein C,Williams AF,Ziegler A.Production of monoclonal antibodies to group A erythrocytes,HLA and other human cell surface antigens-new tools for genetic analysis.Cell.1978 May;14(1):9-20.

[0734] 60.Goldman JM,Hibbin J,Kearney L,Orchard K,Th'ng KH.HLA-DR monoclonal antibodies inhibit the proliferation of normal and chronic granulocytic leukaemia myeloid progenitor cells.Br J Haematol.1982 Nov;52(3):411-20.

[0735] 61.Eng JK,Jahan TA,Hoopmann MR.Comet:an open-source MS/MS sequence database search tool.Proteomics.2013 Jan;13(1):22-4.doi:10.1002/pmic.201200439.Epub 2012 Dec 4.

[0736] 62.Eng JK,Hoopmann MR,Jahan TA,ERertson JD,Noble WS,MacCoss MJ.A deeper look into Comet--implementation and features.J Am Soc Mass Spectrom.2015 Nov;26(11):1865-74.doi:10.1007/s13361-015-1179-x.Epub 2015 Jun 27.

[0737] 63.Lukas Käll,Jesse Canterbury,Jason Weston,William Stafford Noble and Michael J.MacCoss.Semi-supervised learning for peptide identification from shotgun proteomics datasets.Nature Methods 4:923-925,November 2007

[0738] 64.Lukas Käll,John D.Storey,Michael J.MacCoss and William Stafford Noble.Assigning confidence measures to peptides identified by tandem mass spectrometry.Journal of Proteome Research,7(1):29-34,January 2008

[0739] 65.Lukas Käll,John D.Storey and William Stafford Noble.Nonparametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry.Bioinformatics,24(16):i42-i48,August 2008

[0740] 66.Bo Li and C.olin N.Dewey.RSEM:accurate transcript quantification from RNA-Seq data with or without a referenfe genome.BMC Bioinformaties,12:323,August 2011

[0741] 67.Hillary Pearson,Tariq Daouda,Diana Paola Granados,Chantal Durette,Eric Bonneil,Mathieu Courcelles,Anja Rodenbrock,Jean-Philippe Laverduire,Caroline Côté,Sylvie Mader,Sébastien Lemieux,Pierre Thibault,and Claude Perreault.MHC class I-associated peptides derive from selective regions of the human genome.The Journal of Clinical Investigation,2016,

[0742] 68.Juliane Liepe,Fabio Marino,John Sidney,Anita Jeko,Daniel E.Bunting,Alessandro Sette,Peter M.Kloetzel,Michael P.H.Stumpf,Albert J.R.Heck,Michele Mishto.A large fraction of HLA class I ligands are proteasome-generated spliced peptides.Science,21,October 2016.

[0743] 69.Mommen GP.,Marino,F.,Meiring HD.,Poelen,MC.,van Gaans-van den Brink,JA.,Mohammed S.,Heck AJ.,and van Els CA.Sampling From the Proteome to the Human Leukocvte Antigen-DR(HLA-DR)Ligandome Proceeds Via High Specificity.Mol Cell Proteomics 15(4):1412-1423,April 2016

[0744] 70.Sebastian Kreiter,Mathias Vormehr,Niels van de Roemer,Mustafa Diken,Martin Löwer,Jan Diekmann,Sebastian Boegel,Barbara Schrörs,Fulvia Vascotto,John C.Castle,Arbel D.Tadmor,Stephen P.Schoenberger,Christoph Huber,Özlem Türeei,and Ugur Sahin.Mutant MHC class II epitopes drive therapeutic immune responses to caner.Nature 520,692-696,April 2015.71.Tran E.,Turcotte S.,Gros A.,Robbins P.F.,Lu Y.C,Dudley M.E.,Wunderlich J.R.,Somerville R.P.,Hogan K.,Hinrichs C.S.,Parkhurst M.R.,Yang J.C,Rosenberg S.A.Cancer immunotherapy based on mutation-specific CD4+ Tcells in a patient with epithelial cancer.Science 344(6184)641-645,May 2014.72.Andreatta M.,Karosiene E.,Rasmussen M.,Stryhn A.,Buus S.,Nielsen M.Accurate pan-specific prediction of peptide-MHC class II binding affinity with imroved binding core identification.Immunogenetics 67(11-12)641-650,November 2015.

[0745] 73.Nielsen,M.,Lund,O.NN-align.An artificial neural network-based

alignment algorithm for MHC class II peptide binding prediction.BMC Bioinformatics 10:296,September 2009.

[0746] 74.Nielsen,M.,Lundegaard,C.,Lund,O.Prediction of MHC class II binding affinity using SMM-align,a novel stabilization matrix alignment method.BMC Bioinformatics 8:238,July 2007.

[0747] 75.Zhang,J.,et al.PEAKS DB:de novo sequencing assisted database search for sensitive and accurate peptide identification.Molecular&Cellular Proteomics.11(4):1-8.1/2/2012.

[0748] 76.Livingston,B.,et al.A Rational Strategy to Design Multiepitope Immunogens Based on Multiple Th Lymphocyte Epitopes.J.Immunol 168(11)5499-5506,June 2002.

[0749] 77.Timothy O'Donnell,Alex Rubinsteyn,Maria Bonsack,Angelika Riemer,Jeffrey Hammerbacher.MHCflurry:open-source class I MHC binding affinity prediction.https://doi.org/10.1101/174243MHC.https://www.biorxiv.org/content/early/2017/08/09/174243.

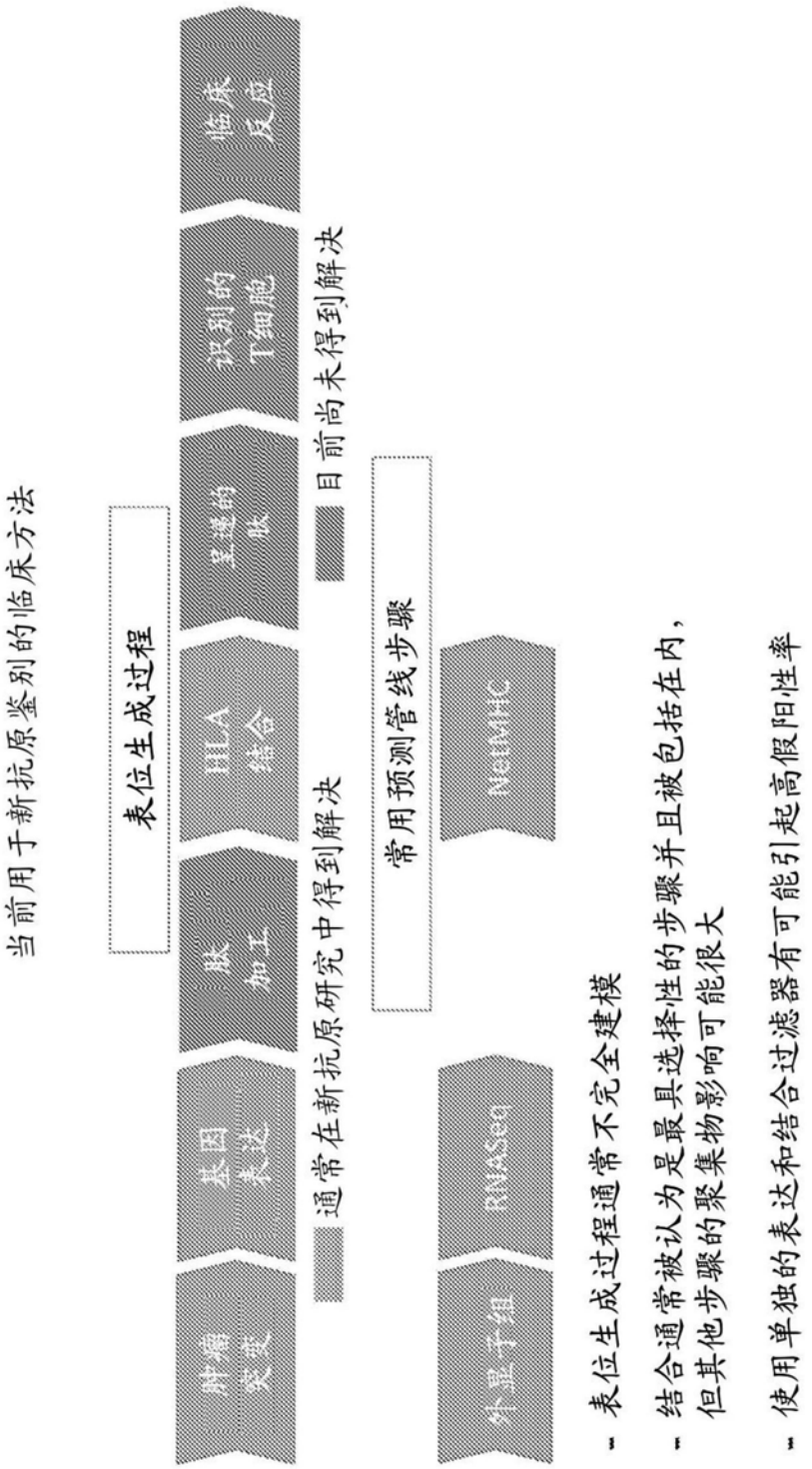
当前用于新抗原鉴别的临床方法
表位生成过程
肿瘤突变
基因表达
肽加工
HLA结合
呈递的肽
识别的T细胞
临床反应
通常在新抗原研究中得到解决
目前尚未得到解决
常用预测管线步骤
外显子组
RNASeq
MuTect
表位生成过程通常不完全建模
结合通常被认为是最具选择性的步骤并且被包括在内，但其他步骤的聚集物影响可能很大
使用单独的表达和结合过滤器有可能引起高假阳性率

图1A

最新文献表明<5%的预测结合肽
可被发现呈递在肿瘤细胞上
外显子组变化
外显子组编码变化
转录物编码变化
预测的新表位
光谱鉴别的新表位
预测的免疫原性新表位
验证的免疫原性新表位
仅7/170 (4%) 的预测新表位
被洗脱并通过质谱法检测到
可能由特定质谱检测方法的
灵敏度限值引起
Yadav, Nature 第515卷 2014年11月17日
NetMHCcon IC50≤500nM
基于MS的鉴别
HLAp
未检测到
结合物
非结合物
1,579
47,777
153
1,140,967
3%
NetMHCcon IC50≤50nM
基于MS的鉴别
HLAp
未检测到
结合物
非结合物
1,133
14,059
599
1,174,685
Bassani-Sternberg M, Mol Cell Proteomics. 2015;14(3):658-73

图1B

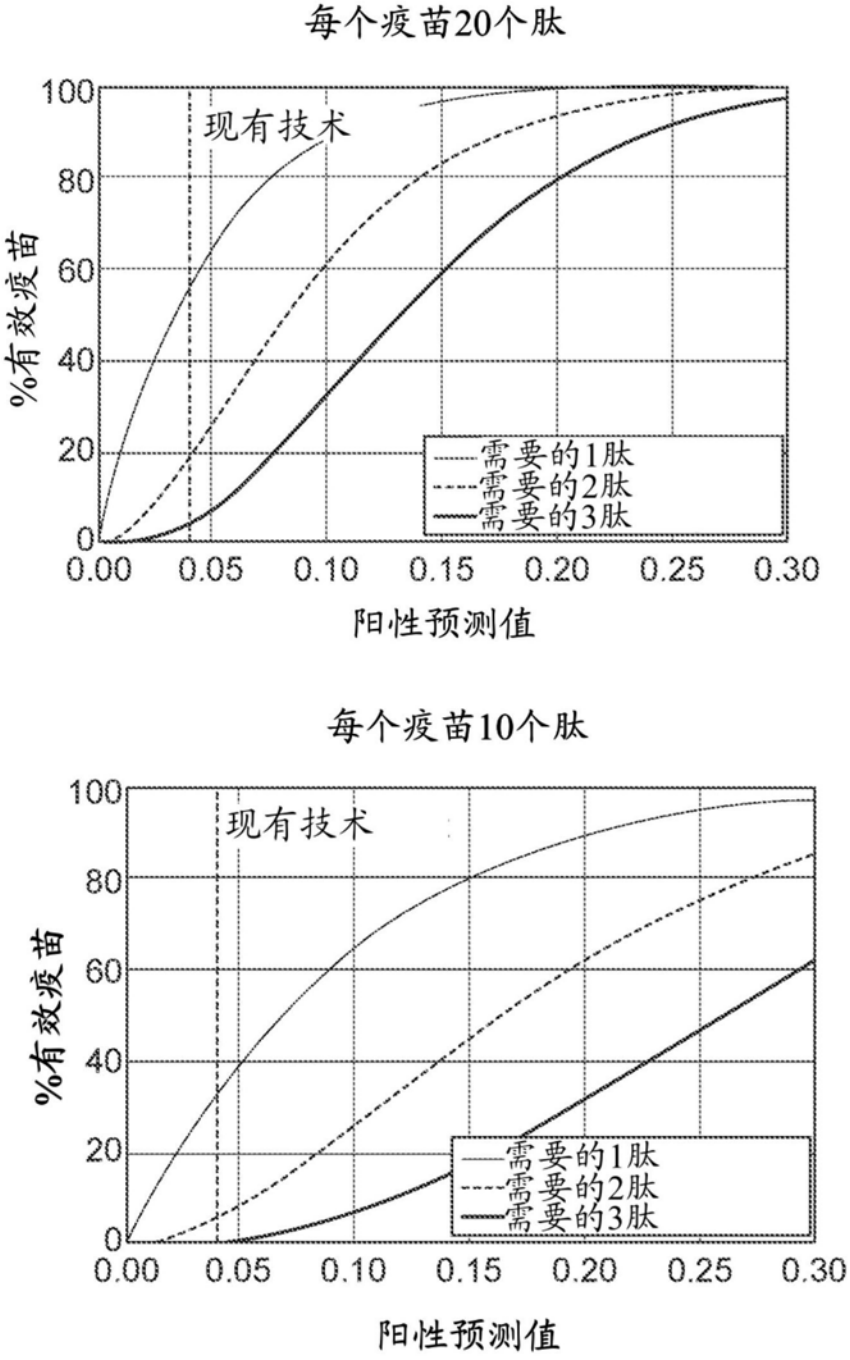
每个疫苗20个肽
现有技术
%有效疫苗
阳性预测值
需要的1肽
需要的2肽
需要的3肽
每个疫苗10个肽
现有技术
%有效疫苗
阳性预测值
需要的1肽
需要的2肽
需要的3肽

图1C

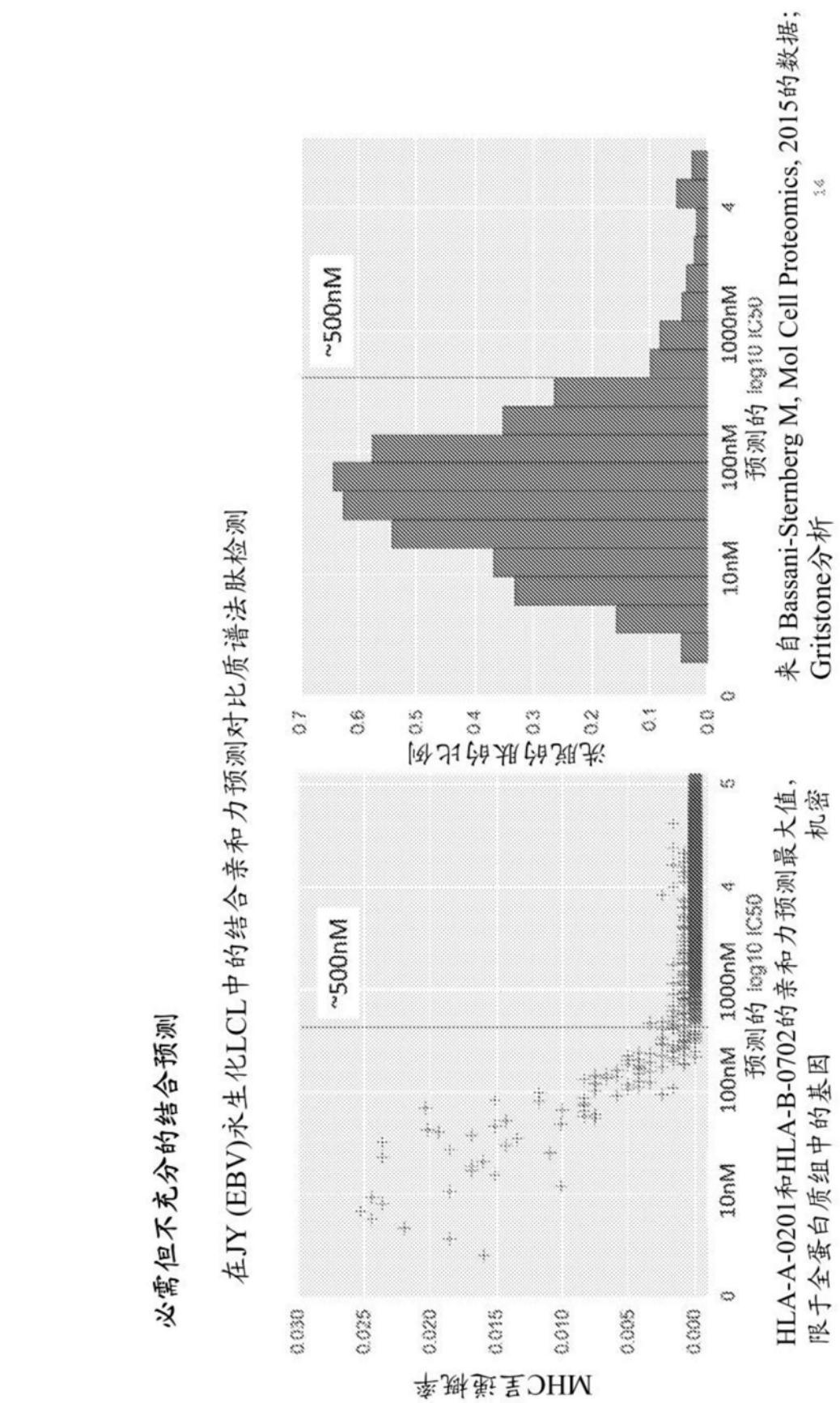
必需但不充分的结合预测
在JY (EBV)永生化LCL中的结合亲和力预测对比质谱法肽检测
~500nM
~500nM
MHC呈递概率
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10nM
100nM
1000nM
4
5
预测的 log10 IC50
来自Bassani-Sternberg M, Mol Cell Proteomics, 2015的数据；
Gritstone分析
HLA-A-0201和HLA-B-0702的亲和力预测最大值，
限于全蛋白质组中的基因
机密

图1D

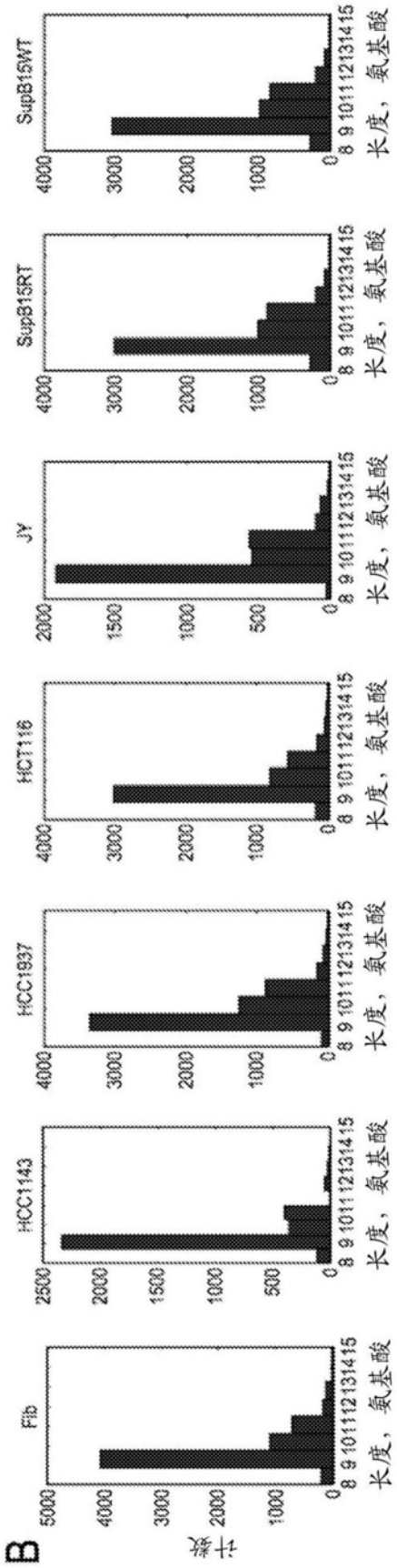

图1E

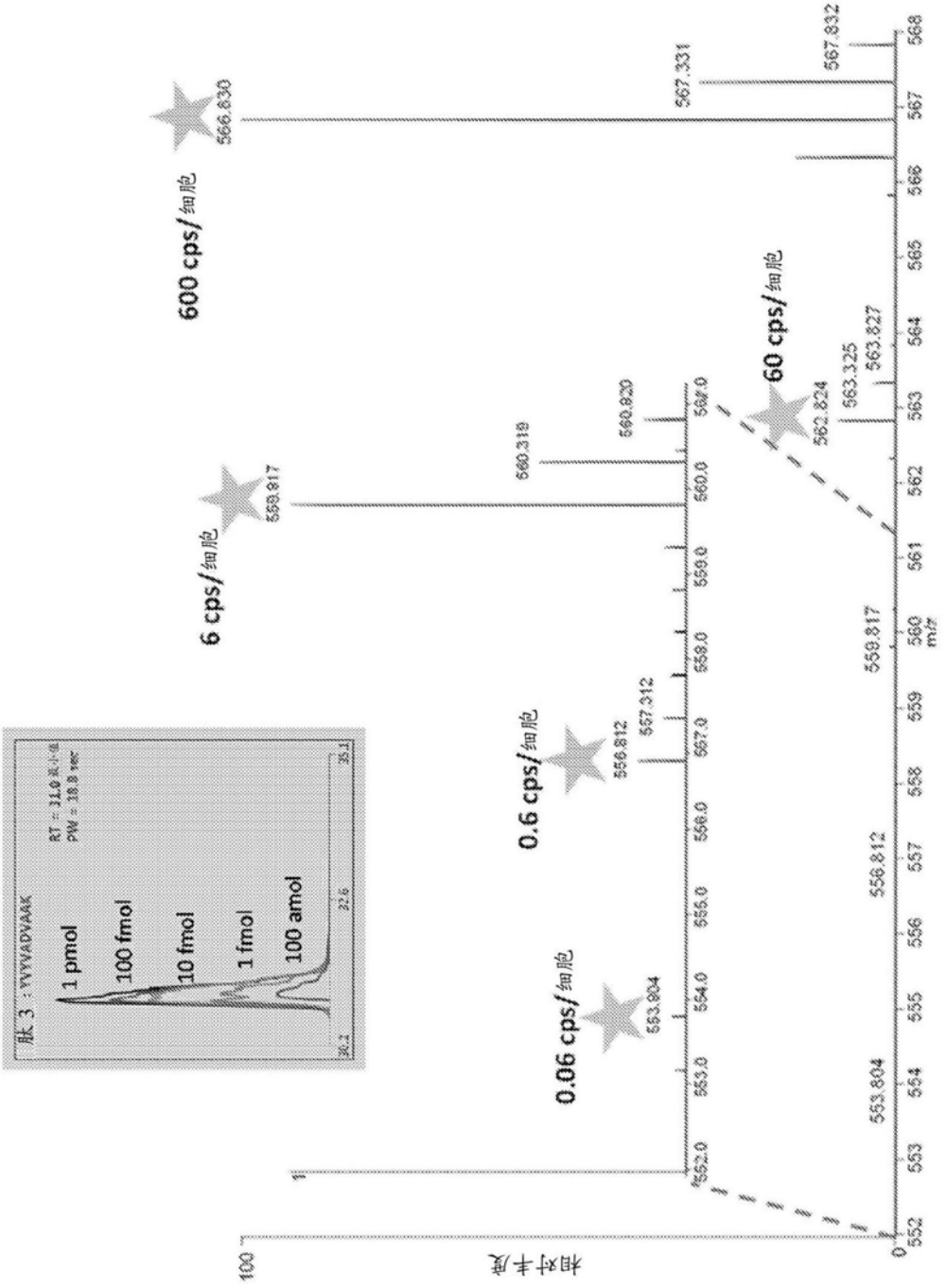
肽 3 : YVYVADVAAK
RT = 31.0
PW = 38.8 sec
1 pmol
100 fmol
10 fmol
1 fmol
100 amol
600 cps/细胞
6 cps/细胞
0.6 cps/细胞
0.06 cps/细胞
60 cps/细胞
相对丰度
m/z

图1F

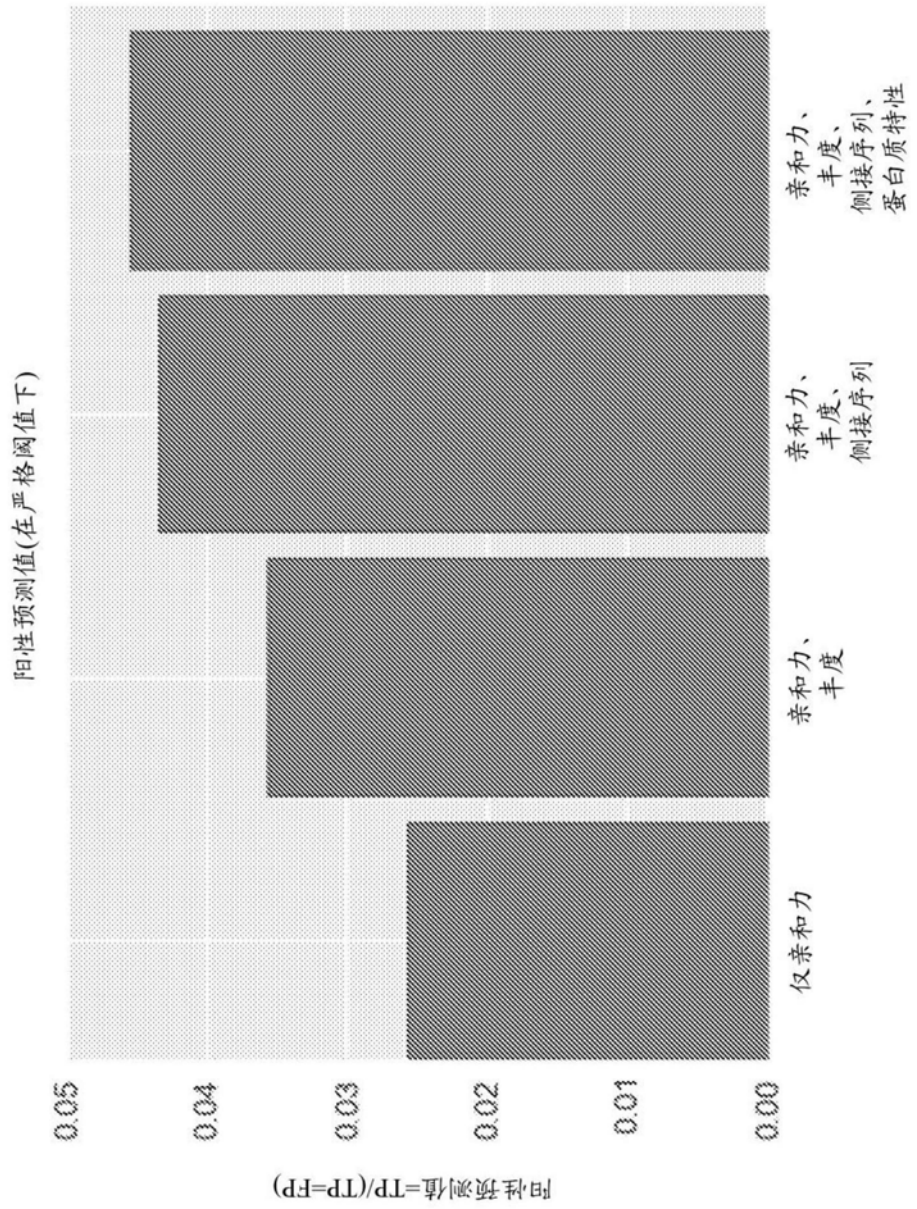
阳性预测值(在严格阈值下)
0.05
0.04
0.03
0.02
0.01
0.00
仅亲和力
亲和力、丰度
亲和力、丰度、侧接序列
亲和力、丰度、侧接序列、蛋白质特性
阳性预测值=TP/(TP=FP)

图1G

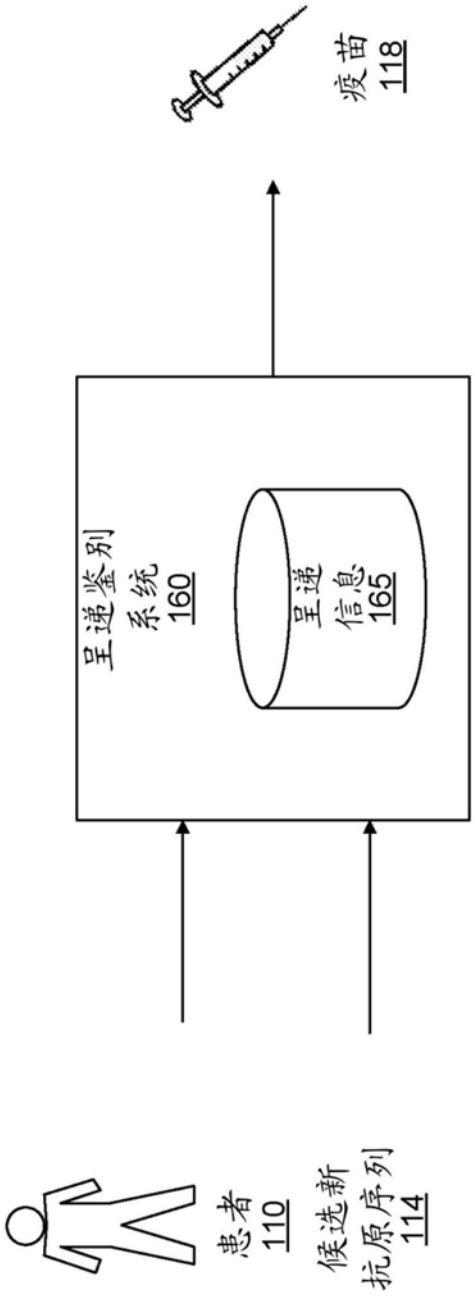

图2A

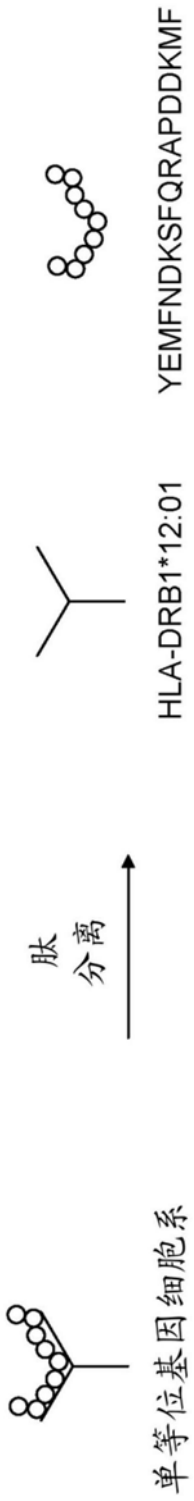

图2B

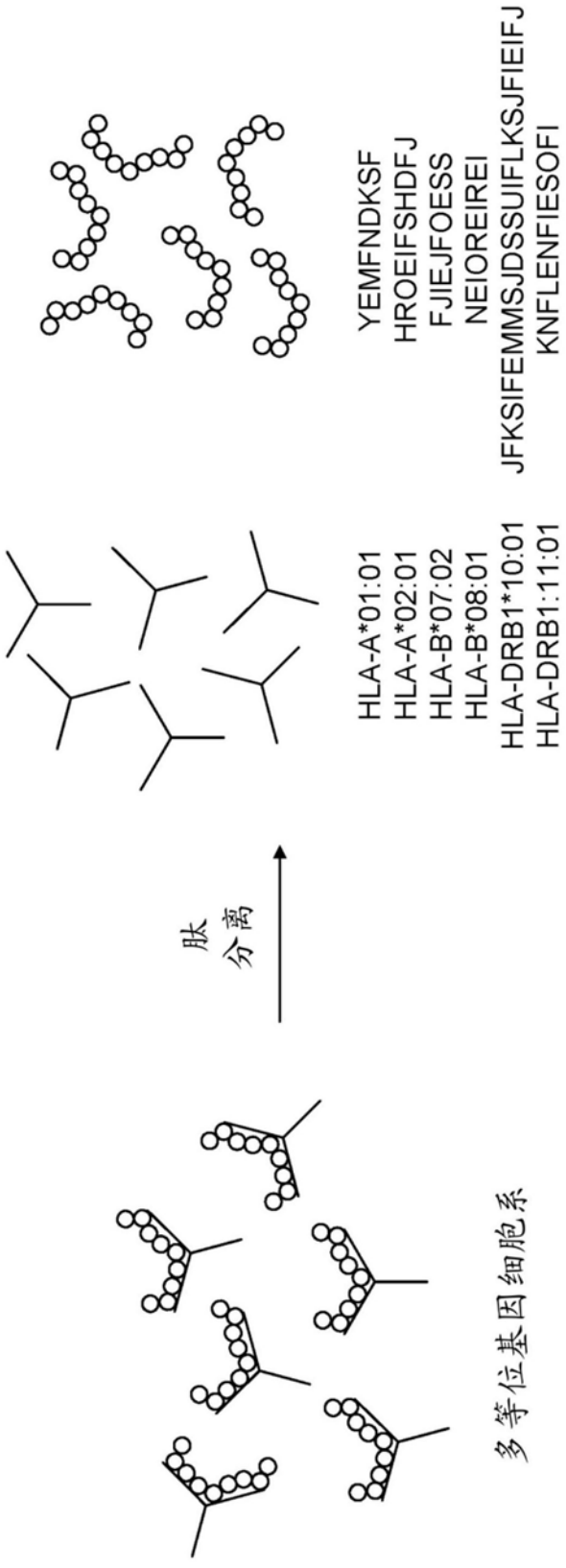
多等位基因细胞系
肽
分离
HLA-A*01:01
HLA-A*02:01
HLA-B*07:02
HLA-B*08:01
HLA-DRB1*10:01
HLA-DRB1:11:01
YEMFNDKSF
HROEIFSHDFJ
FJIEJFOESS
NEIOREIREI
JFKSIFEMMSJDSSUIFLKSJFIEIFJ
KNFLENFIESOFI

图2C

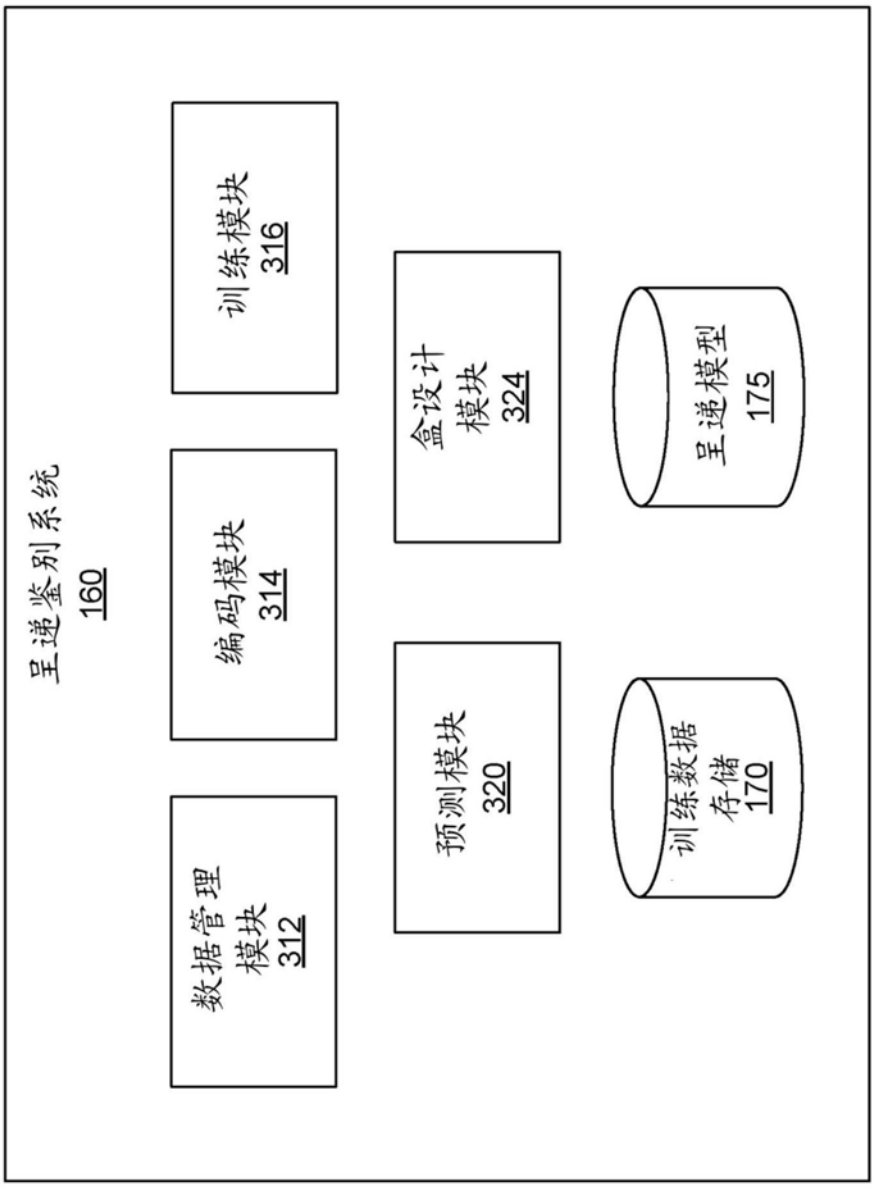
呈递鉴别系统
160
数据管理模块
312
编码模块
314
训练模块
316
预测模块
320
盒设计模块
324
训练数据存储
170
呈递模型
175

图3

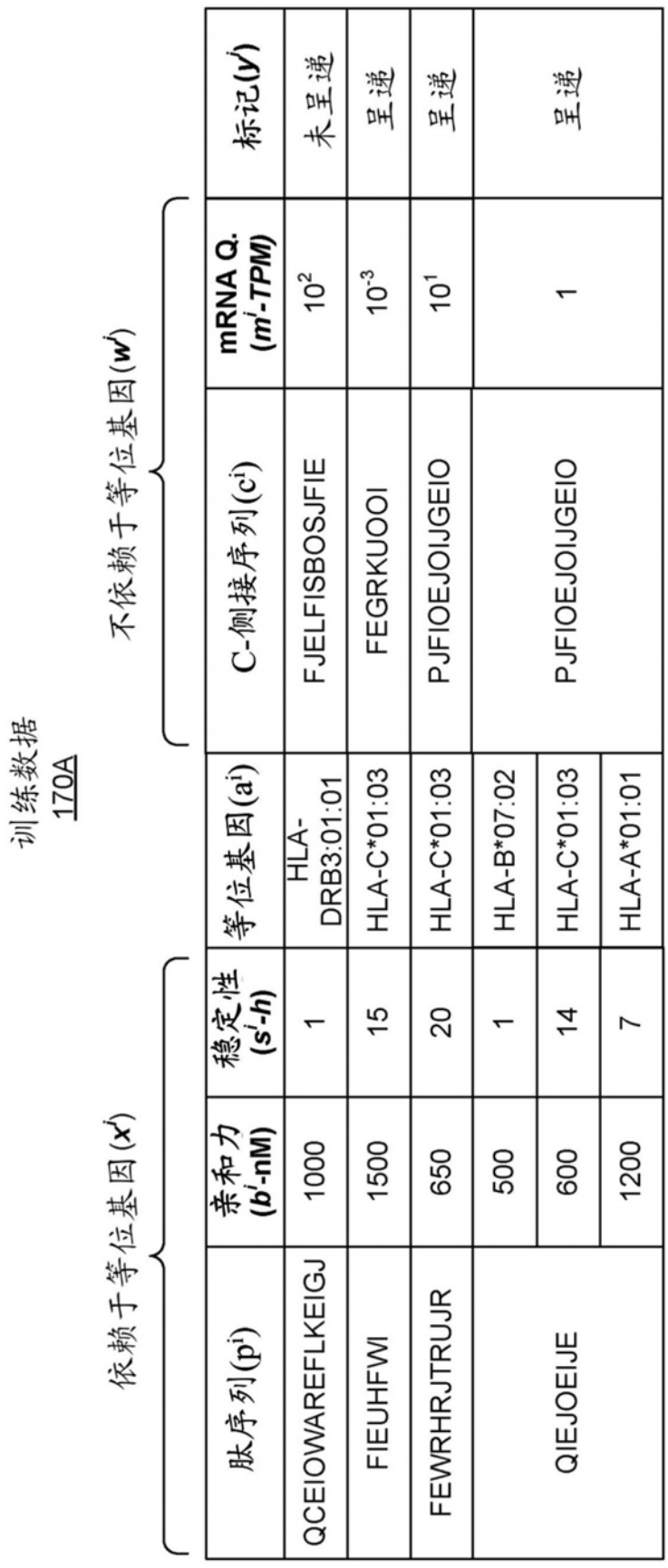
训练数据
170A
依赖于等位基因(x^i)
不依赖于等位基因(w^i)
肽序列(p^i)
亲和力 (b^i-nM)
稳定性 (s^i-h)
等位基因(a^i)
C-侧接序列(c^i)
mRNA Q. (m^i-TPM)
标记(y^i)
QCEIOWAREFLKEIGJ
1000
1
HLA-DRB3:01:01
FJELFISBOSJFIE
10^2
未呈递
FIEUHFWI
1500
15
HLA-C*01:03
FEGRKUOOI
10^{-3}
呈递
FEWRHRJTRUJR
650
20
HLA-C*01:03
PJFIOEJOIJGEIO
10^1
呈递
QIEJOEIJE
500
1
HLA-B*07:02
PJFIOEJOIJGEIO
1
呈递
600
14
HLA-C*01:03
1200
7
HLA-A*01:01

图4

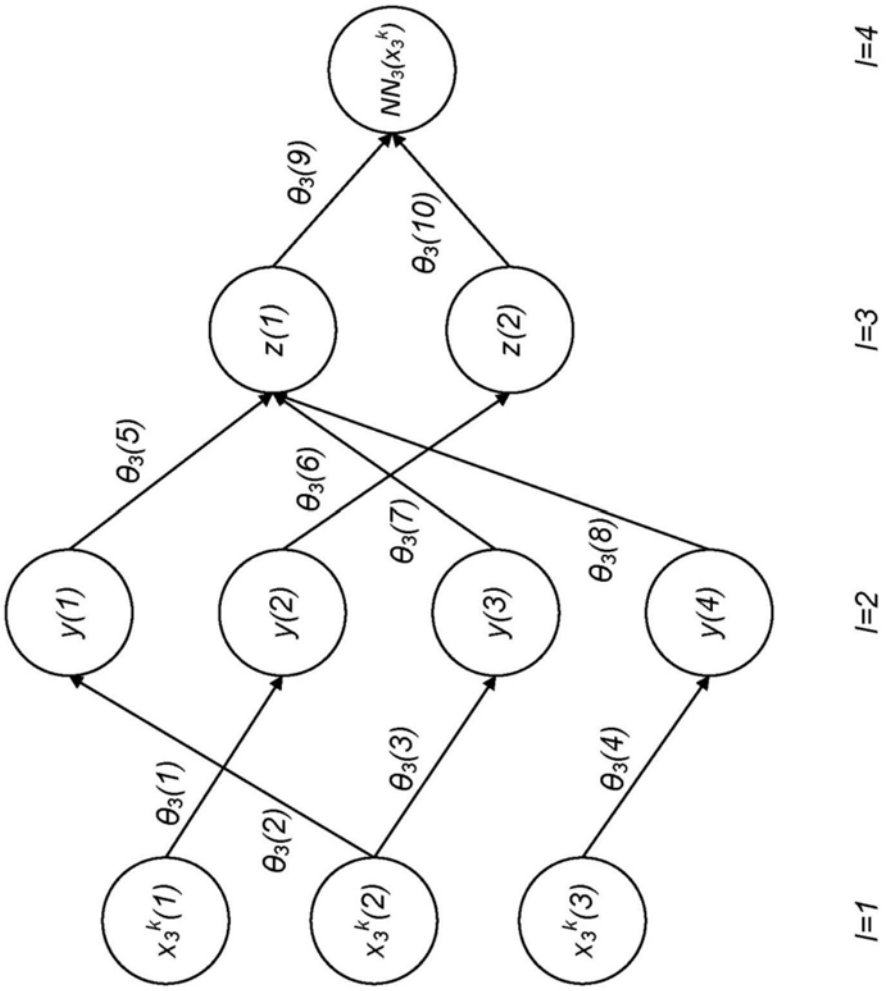
$NN_3(x_3^k)$
$\theta_3(9)$
$\theta_3(10)$
z(1)
z(2)
$\theta_3(5)$
$\theta_3(6)$
$\theta_3(7)$
$\theta_3(8)$
y(1)
y(2)
y(3)
y(4)
$\theta_3(1)$
$\theta_3(2)$
$\theta_3(3)$
$\theta_3(4)$
$x_3^k(1)$
$x_3^k(2)$
$x_3^k(3)$
l=4
l=3
l=2
l=1

图5

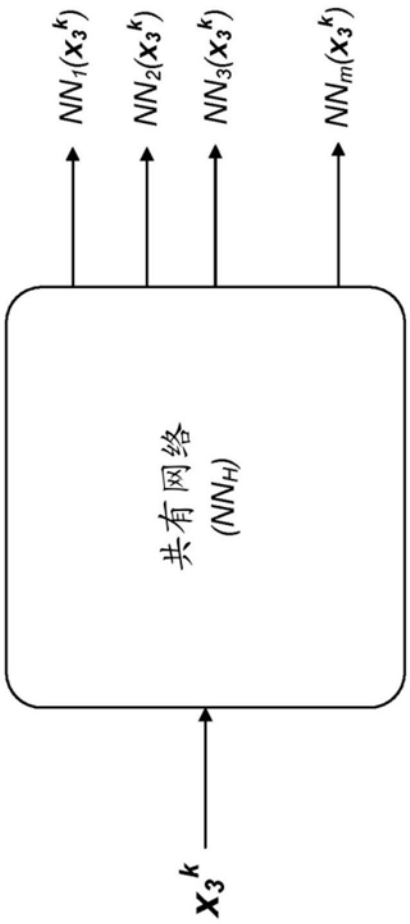

图6A

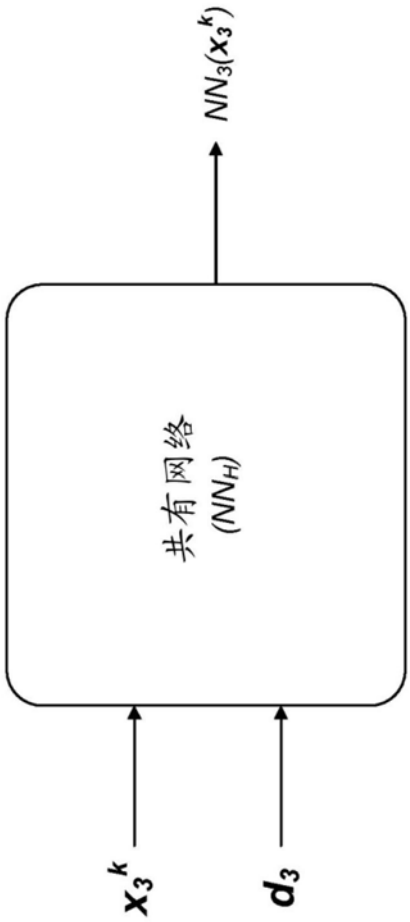

图6B

图7

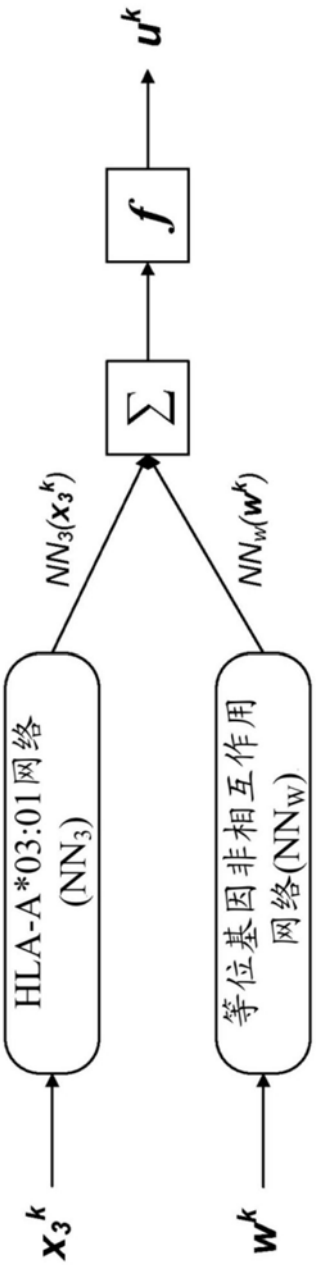

图8

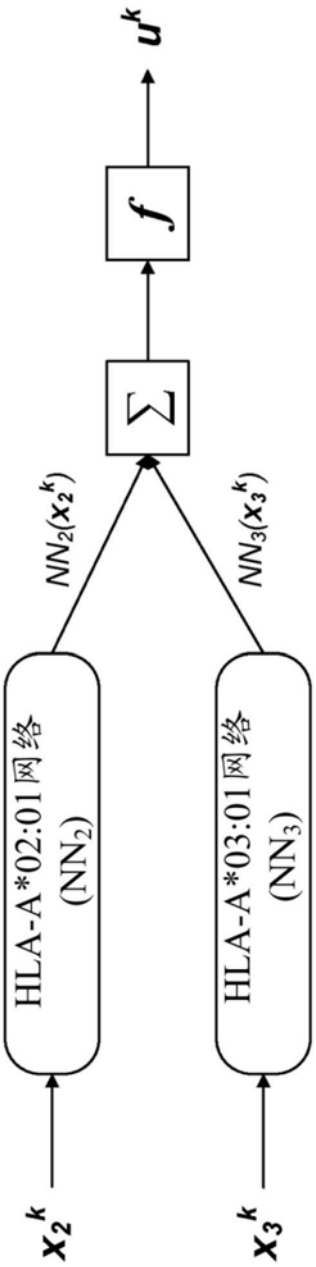
u^k
f
$\sum$
$NN_2(x_2^k)$
$NN_3(x_3^k)$
HLA-A*02:01网络
(NN_2)
HLA-A*03:01网络
(NN_3)
x_2^k
x_3^k

图9

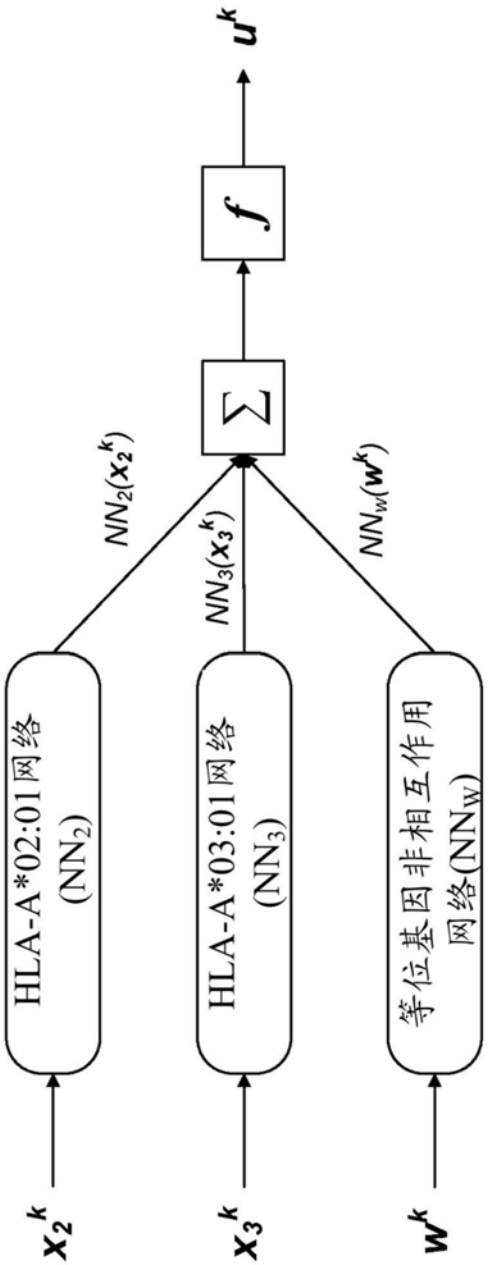
u^k
f
$\sum$
$NN_2(x_2^k)$
$NN_3(x_3^k)$
$NN_w(w^k)$
HLA-A*02:01网络 (NN_2)
HLA-A*03:01网络 (NN_3)
等位基因非相互作用网络(NN_W)
x_2^k
x_3^k
w^k

图10

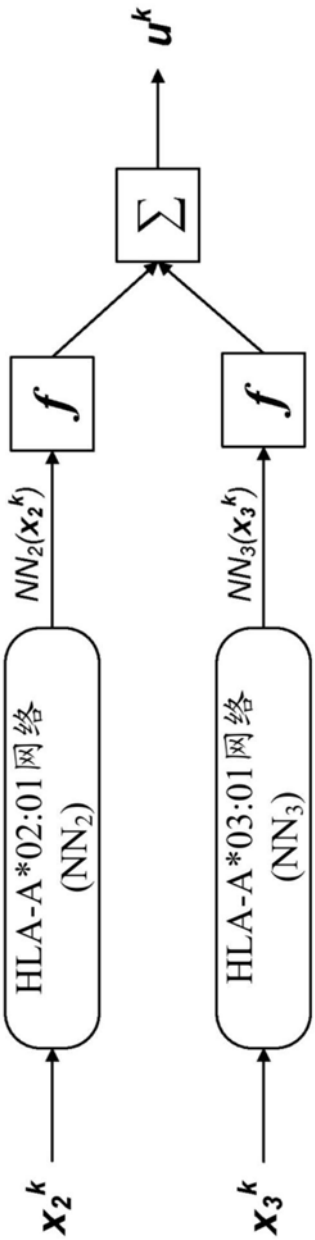
u^k
$\sum$
f
f
$NN_2(x_2^k)$
$NN_3(x_3^k)$
HLA-A*02:01网络
(NN_2)
HLA-A*03:01网络
(NN_3)
x_2^k
x_3^k

图11

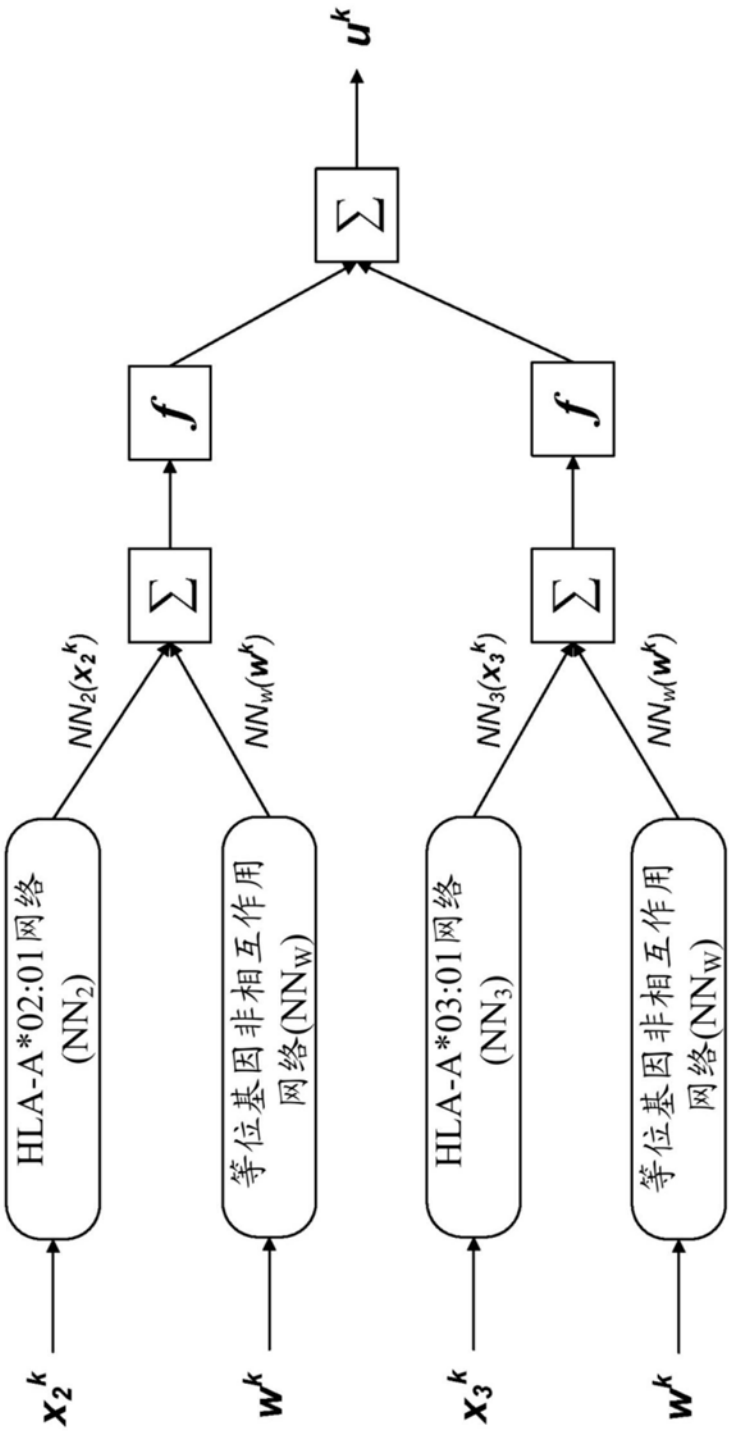
u^k
∑
f
f
∑
∑
$NN_2(x_2^k)$
$NN_w(w^k)$
$NN_3(x_3^k)$
$NN_w(w^k)$
HLA-A*02:01网络 (NN2)
等位基因非相互作用网络(NNW)
HLA-A*03:01网络 (NN3)
等位基因非相互作用网络(NNW)
x_2^k
w^k
x_3^k
w^k

图12

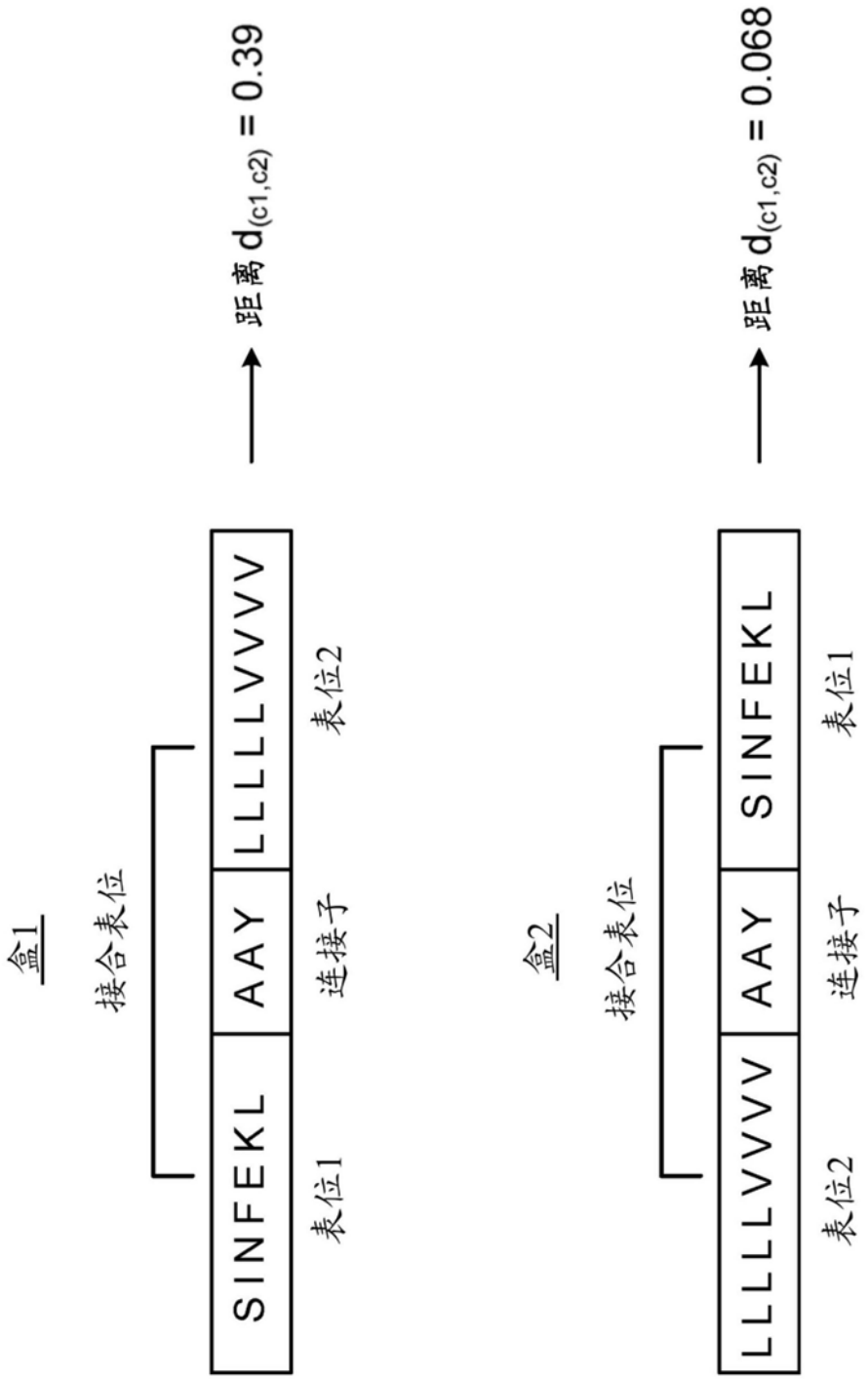
盒1
接合表位
SINFEKL
AAY
LLLLLVVVV
表位1
连接子
表位2
距离 $d_{(c1,c2)}$ = 0.39
盒2
接合表位
LLLLLVVVV
AAY
SINFEKL
表位2
连接子
表位1
距离 $d_{(c1,c2)}$ = 0.068

图13

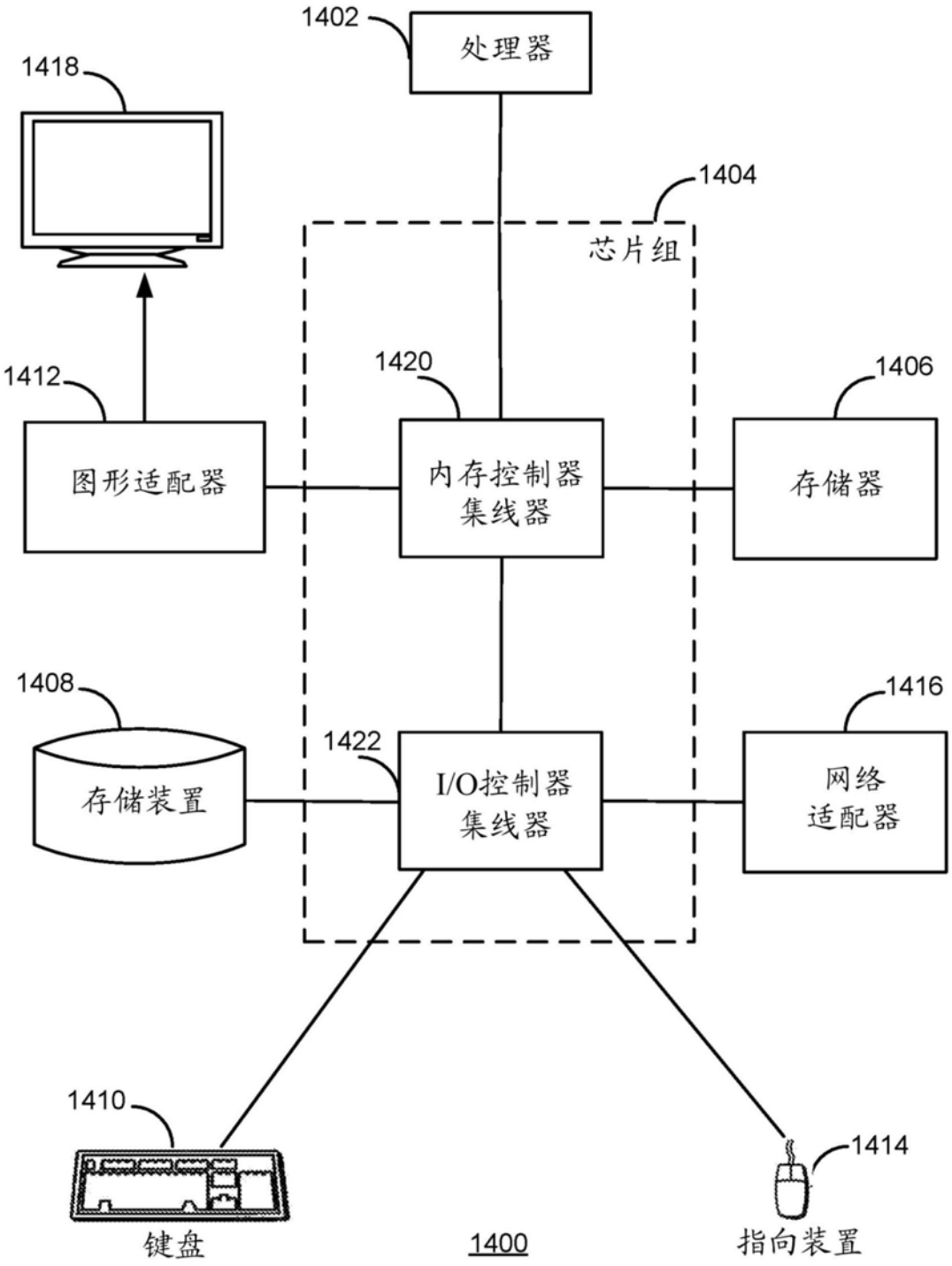
1402
处理器
1418
1404
芯片组
1412
1420
1406
图形适配器
内存控制器
集线器
存储器
1408
1416
1422
存储装置
I/O控制器
集线器
网络
适配器
1410
1414
键盘
1400
指向装置

图14