(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113762416 B
(45)授权公告日 2023.05.30

(21)申请号 202111204465.5

(22)申请日 2021.10.15

(65)同一申请的已公布的文献号
申请公布号 CN 113762416 A

(43)申请公布日 2021.12.07

(73)专利权人 南京澄实生物科技有限公司
地址 210000 江苏省南京市江北新区探秘路73号树屋十六栋D-2栋2层209室

(72)发明人 费才溢 徐实

(74)专利代理机构 南京天华专利代理有限责任公司 32218
专利代理师 刘畅 傅婷婷

(51)Int.Cl.
G16B 15/30(2019.01)
G06F 18/25(2023.01)
G06F 18/2135(2023.01)
G06F 18/214(2023.01)
G06N 3/04(2023.01)
G06N 3/08(2023.01)
G16B 30/10(2019.01)

(56)对比文件
CN 111105843 A,2020.05.05
CN 113139568 A,2021.07.20
US 2017262984 A1,2017.09.14
方春;孙福振;李彩虹;宋莉.基于长短期记忆网络的抗癌肽的预测.山东理工大学学报(自然科学版).2020,(03),37-42.
BinBin Chen et al..Predicting HLA class II antigen presentation through integrated deep learning.《Nature Biotechnology》.2019,1332-1343.

审查员 黄沛

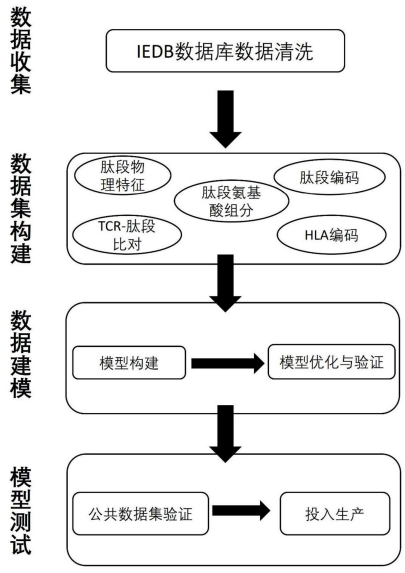
权利要求书3页 说明书9页 附图3页

(54)发明名称

基于多模态深度编码的抗原免疫原性预测方法和系统

(57)摘要

本发明公开了一种基于多模态深度编码的抗原免疫原性预测方法和系统,包括:1)包含较全面HLA等位基因个数;2)可变长肽段编码方式,涵盖主要HLA-I结合肽段长度;3)考虑抗原受体谱系对肽段免疫原性的影响;4)包含肽段和MHC序列的物理和氨基酸组成特征;5)多模态特征融合得到预测分数并进行预测。不同于以往的仅基于生物实验或较单一化的数据模态的预测模型,本系统方案能高效地融合多模态信息,进行更加准确高效的预测。基于真实数据结果表面,其TOP-10结果具有较高水平的PPV值,能够更好的应用在真实的药物研发生产环境。



CN 113762416 B

1.一种基于多模态深度编码的抗原免疫原性预测方法,其特征在于它包括以下步骤:

S1、特征选择:选定与抗原免疫原性相关的特征,作为待融合特征;待融合特征选择为:肽段序列特征、HLA-I序列特征、抗原受体-肽段互作特征、肽段物理特征、肽段氨基酸组分特征;

S2、归一化处理:设置隐嵌入维度作为不同待融合特征的最终输入维度,将S1中获取的待融合特征进行变换和尺度缩放,获得标准特征;根据S1中特征选择的不同,S2中选定相应的归一化处理方案,以获取格式、维度统一,便于融合的标准特征,具体为:

-序列特征:使用主成分分解PCA对AAindex数据库中的AAindex1数据进行变换,选取变换后的12个主成分特征,为输入的序列进行编码作为序列特征的标准特征;

-互作特征:基于AAindex数据库中的AAindex3矩阵,进行序列比对,获取比对分数,通过尺度缩放获取互作特征的标准特征;

-物理特征:表示序列的电荷、疏水性、不稳定性特征,通过尺度缩放获得物理特征的标准特征;

-组分特征:表示序列氨基酸组分特征,统计其标准氨基酸编码出现的次数作为组分特征的标准特征;

S3、特征融合:将维度相同的标准特征作线性融合操作,融合后的特征向量/矩阵输入深度神经网络,进行非线性变换与融合,获得抗原免疫原性的最终特征分数;

S4、构建预测模型:特征融合,构建包含最终特征分数的预测模型和优化模型;

S5、求解优化模型,获得最优参数的预测模型;

S6、使用最优参数的预测模型进行抗原免疫原性预测。

2.根据权利要求1所述的方法,其特征在于S2中:

肽段序列特征通过以下方法获得标准特征:采用主成分分解PCA对AAindex数据库中的AAindex1特征进行变换,选取变换后的12个主成分特征,对肽段的蛋白组成氨基酸进行编码作为肽段序列特征;

HLA-I序列特征通过以下方法获得标准特征:采用主成分分解PCA对AAindex数据库中的AAindex1特征进行变换,选取变换后的12个主成分特征,对HLA-I序列的蛋白组成氨基酸进行编码作为HLA-I序列特征;

抗原受体-肽段互作特征通过以下方法获得标准特征:基于AAindex数据库中的AAindex3特征,进行序列比对,获取比对分数,通过尺度缩放获得标准化特征,作为抗原受体-肽段互作特征;

肽段物理特征通过尺度缩放获得标准特征,以保证模型训练优化过程的数值稳定性;

肽段氨基酸组分特征通过以下方法获得标准特征:统计肽段中标准氨基酸编码出现的次数,作为肽段氨基酸组分特征。

3.根据权利要求2所述的方法,其特征在于抗原受体-肽段互作特征的尺度变化公式为:

$$p' = \frac{p - \min(p)}{\max(p) - \min(p)}$$

式中,p表示抗原受体-肽段互作分数,p'表示作为标准特征的抗原受体-肽段互作分数。

4.根据权利要求2所述的方法,其特征在于肽段物理特征的尺度变化公式为:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

式中，x表示肽段物理特征分数，x'表示作为标准特征的肽段物理特征分数。

5.根据权利要求1所述的方法，其特征在于S3特征融合中，所述线性融合操作包括点乘、或加和、或组合成特征矩阵。

6.根据权利要求1所述的方法，其特征在于S4中构建优化模型：

$$min\frac{\sum_{n=1}^{N}\left(y_n - f_W(x_n)\right)^2}{N}$$

式中，f是含可学习参数的预测模型；W表示该模型中可学习参数，包括各融合特征获取时方案权重；x_n表示输入的特定数据，y_n是训练数据中免疫原性可信度数值；N表示样本总数。

7.根据权利要求1所述的方法，其特征在于S5中求解优化模型：多次遍历所有训练数据，利用基于随机梯度优化方法的优化器进行优化，得到最优的模型参数，获得预测模型f_W。

8.根据权利要求7所述的方法，其特征在于采用Beta分布，对IEDB数据库中具有实验验证的免疫原性结果进行编码，转换为回归拟合问题，基于此来提高训练模型可靠性；所述训练数据经过包括重抽样、剔除负样本的数据处理，避免使用的训练的数据正负样本量不平衡的问题。

9.一种基于多模态深度编码的抗原免疫原性预测系统，其特征在于它包括：

-数据收集模块，整理IEDB数据库中具有免疫原性验证结果的肽段及其MHC-I配体数据对；

-数据集构建，根据整理的IEDB数据库中具有免疫原性验证结果的肽段及其MHC-I配体数据对，构建正负样本；

-数据建模模块，构建预测模型并求解预测模型；

所述数据建模模块包括：

-特征选择模块，选定与肽段免疫原性相关的特征，作为待融合特征；待融合特征选择为：肽段序列特征、HLA-I序列特征、抗原受体-肽段互作特征、肽段物理特征、肽段氨基酸组分特征；

-归一化处理模块，将待融合特征进行变换和尺度缩放，获得标准特征；

-特征融合模块，将多模态的标准特征输入深度神经网络进行融合，得到肽段免疫原性最终特征分数；标准特征，具体为：

-序列特征；使用主成分分解PCA对AAindex数据库中的AAindex1数据进行变换，选取变换后的12个主成分特征，为输入的序列进行编码作为序列特征的标准特征；

-互作特征：基于AAindex数据库中的AAindex3矩阵，进行序列比对，获取比对分数，通过尺度缩放获取互作特征的标准特征；

-物理特征：表示序列的电荷、疏水性、不稳定性特征，通过尺度缩放获得物理特征的标准特征；

-组分特征：表示序列氨基酸组分特征，统计其标准氨基酸编码出现的次数作为组分特

征的标准特征；

-预测模型构建模块，构建包含最终特征分数的预测模型和优化模型；

-预测模型求解模块，计算获得最优参数的预测模型。

10.根据权利要求9所述的系统，其特征在于所述特征融合模块中，采用可变长的维度的输入设计，以便未来加入更多新模态特征的接口。

11.根据权利要求9所述的系统，其特征在于采用Beta分布，对IEDB数据库中具有实验验证的免疫原性结果进行编码，转换为回归拟合问题，基于此来提高训练模型可靠性；所述训练数据经过包括重抽样、剔除负样本的数据处理，避免使用的训练的数据正负样本量不平衡的问题。

12.根据权利要求9所述的系统，其特征在于它还包括：

-测试模块，整理包含在文献中但不包含在数据集构建模块出现的数据，以待优化得到最优模型后，验证测试模型的对于未曾见过的免疫原性数据对的预测效果。

基于多模态深度编码的抗原免疫原性预测方法和系统

技术领域

[0001] 本发明涉及生物信息学领域,尤其涉及一种基于深度编码与多模态融合的预测新生抗原免疫原性的方法和系统。

背景技术

[0002] 免疫治疗已成为一种很有希望的癌症治疗策略。各种形式的免疫治疗可以增强免疫系统以抵抗癌症,或者使免疫系统更容易识别并摧毁癌细胞,或减慢其生长。有效的靶向免疫治疗需要精确的预测哪些癌症特异性新肽段最有可能引起免疫反应。

[0003] CD8+T细胞免疫反应是识别和杀死感染细胞和恶性肿瘤细胞的关键。过去的十年中癌症免疫治疗表明,利用增强CD8+T细胞的介导对癌细胞的控制和清除具有临床意义。在分子水平层面,CD8+T细胞对肽表位的识别基于一系列特定事件。首先,肽段被蛋白酶从源蛋白上切割,转运到内质网中并与HLA-I分子结合。稳定结合后,肽-HLA-I(pHLA)复合物被呈现在细胞表面。随后,T细胞受体(TCR)可以与pHLA复合物结合,从而启动免疫突触的形成,并最终导致被感染或恶性细胞的死亡。

[0004] 基于HLA呈递-CD8+T细胞识别原理的癌症疫苗是当今医学与药物学的热点问题。肿瘤疫苗教导免疫系统将传染性病原体或癌细胞识别为需要消除的外来物质。癌细胞表面存在特殊的蛋白质,通过靶向这些蛋白质,免疫系统可以特异性地消除癌细胞,同时不伤害正常的细胞。此外,疫苗还能防止癌症复发,清除治疗后残留的癌细胞。肿瘤疫苗的分类方法有很多种,依据治疗原理可以划分为预防性和治疗性疫苗两大类,治疗性肿瘤疫苗还可以依据靶点类型和疫苗成药载体的不同进行划分。

[0005] 其中以mRNA作为载体的治疗性肿瘤疫苗有以下几点突出优势:(1)mRNA可以同时编码多种抗原,具有MHCI和MHCII结合表位的完整蛋白质,以促进体液和细胞适应性免疫反应,提供更强化的抗肿瘤免疫力。(2)与DNA疫苗相比,mRNA疫苗是非整合的,高度可降解的,没有插入诱变潜力。(3)与蛋白质或细胞介导的疫苗相比,mRNA的IVT产生不含细胞和致病性病毒成分,没有感染可能性,正在进行临床试验测试的大多数mRNA疫苗通常具有良好的耐受性,罕有注射部位反应。(4)mRNA癌症疫苗的另一个优点是快速和可扩展的制造。

[0006] 随着两种用于预防COVID-19的mRNA-LNP疫苗获得批准,mRNA技术路线的可行性和优势已经得到了广泛的认可,并且随着资本的关注以及越来越多的研究人员的参与,mRNA疫苗乃至mRNA药物开发正在经历相当大的爆发式发展。其中一个关键的核心技术点,就是预测mRNA疫苗的核心靶标:新生抗原(Neoantigen)的肿瘤特异性抗原TSA。Neoantigen来源于肿瘤细胞中的随机体细胞突变,不存在于正常细胞中。Neoantigen可被宿主免疫系统识别为“非自身”的序列,引发强烈的免疫反应。预测个性化HLA新生抗原(Neoantigen)疫苗的主要步骤如下:

[0007] (1)鉴定和确认在患者肿瘤中表达的特异性免疫原性非同义体细胞突变。对肿瘤组织进行活组织检查以进行全外显子组或转录组测序。可以通过比较肿瘤和匹配的健康组织的序列来鉴定肿瘤的非同义体细胞突变,例如点突变和插入缺失。

[0008] (2)使用主要组织相容性复合物(MHC)I和II类表位预测算法,分析和鉴定具有最高免疫原性的突变。

[0009] (3)基于体外结合测定结果进一步证实候选抗原的排序列表。

[0010] 步骤(2)中,为了准确预测新生抗原免疫原性,需要知道1)哪些肽段会与MHC结合2)哪些pHLA能够引起免疫反应。当前已经开发了大量的HLA-肽段结合预测工具来预测哪些肽段将与特定的NHC进行结合。然而,仅凭MHC结合预测不足以推断免疫原性,因为此类工具无法预测哪些肽将触发T细胞反应。当前Neoantigen疫苗开发的难点之一是为什么某些被感染或癌症特异性表达的并且被HLA-I呈递到细胞表面的肽段能给被CD8+T细胞识别并引发特定的免疫反应,而有些则不能。随着人工智能技术在生物信息学中的广泛应用,已经有领域内学者开始尝试利用数据驱动的机器学习方法,填补这一空缺。

[0011] 其中代表性的工作与技术有,韩国延世大学团队的工作(参考文献:Kim S,Kim HS,Kim E,et al.Neopepsee:accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information.Ann Oncol.2018;29(4):1030-1036.doi:10.1093/annonc/mdy022)基于14个独立特征开发的机器学习算法来预测肽段免疫原性。拉霍拉学院团队的工作(参考文献:Vita R,Mahajan S,Overton JA,et al.The Immune Epitope Database(IEDB):2018update.Nucleic Acids Res.2019;47(D1):D339-D343.doi:10.1093/nar/gky1006)通过考虑Kullback-Leibler散度和氨基酸偏好的位置加权计算模式来预测肽段免疫原性。浙江大学团队的工作(参考文献:Wu J,Wang W,Zhang J,et al.DeepHLApan:A Deep Learning Approach for Neoantigen Prediction Considering Both HLA-Peptide Binding and Immunogenicity.Front Immunol.2019;10:2559.Published 2019Nov 1.doi:10.3389/fimmu.2019.02559)基于IEDB数据,采用深度学习算法,来预测肽段免疫原性。

[0012] 以上提到的现有主流pHLA新生抗原免疫原性预测方法,具有以下局限性:1)只考虑有限的HLA等位基因个数;2)未考虑不同受试者抗原受体谱系对肽段免疫原性的影响;3)这些算法基本未考虑肽段和MHC氨基酸序列的物理和组成特征。这些方法输出的结果,可能无法完全反应触发T细胞反应的pHLA特征;4)较单一化的数据模态与数据模型,使其无法充分利用生物信息大数据所蕴含的多模态信息,并缺少可扩展性。

发明内容

[0013] 本发明针对背景技术中存在的问题,提出了一种基于多模态深度编码的抗原免疫原性预测方法和系统。

[0014] 技术方案:

[0015] 本发明首先公开了一种基于多模态深度编码的抗原免疫原性预测方法,它包括以下步骤:

[0016] S1、特征选择:选定与抗原免疫原性相关的特征,作为待融合特征;

[0017] S2、归一化处理:设置隐嵌入维度作为不同待融合特征的最终输入维度,将S1中获取的待融合特征进行变换和尺度缩放,获得标准特征;

[0018] S3、特征融合:将维度相同的标准特征作线性融合操作,融合后的特征向量/矩阵输入深度神经网络,进行非线性变换与融合,获得抗原免疫原性的最终特征分数;

[0019] S4、构建预测模型:特征融合,构建包含最终特征分数的预测模型和优化模型;

[0020] S5、求解优化模型,获得最优参数的预测模型;

[0021] S6、使用最优参数的预测模型进行抗原免疫原性预测。

[0022] 优选的,根据S1中特征选择的不同,S2中选定相应的归一化处理方案,以获取格式、维度统一,便于融合的标准特征,具体为:

[0023] -序列特征:使用主成分分解PCA对AAindex数据库中的AAindex1数据进行变换,选取变换后的12个主成分特征,为输入的序列进行编码作为序列特征的标准特征;

[0024] -互作特征:基于AAindex数据库中的AAindex3矩阵,进行序列比对,获取比对分数,通过尺度缩放获取互作特征的标准特征;

[0025] -物理特征:表示序列的电荷、疏水性、不稳定性特征,通过尺度缩放获得物理特征的标准特征;

[0026] -组分特征:表示序列氨基酸组分特征,统计其标准氨基酸编码出现的次数作为组分特征的标准特征。

[0027] 具体的,S1中待融合特征选择为:肽段序列特征、HLA-I序列特征、抗原受体-肽段互作特征、肽段物理特征、肽段氨基酸组分特征。

[0028] 具体的,S2中:

[0029] 肽段序列特征通过以下方法获得标准特征:采用主成分分解PCA对AAindex数据库中的AAindex1特征进行变换,选取变换后的12个主成分特征,对肽段的蛋白组成氨基酸进行编码作为肽段序列特征;

[0030] HLA-I序列特征通过以下方法获得标准特征:采用主成分分解PCA对AAindex数据库中的AAindex1特征进行变换,选取变换后的12个主成分特征,对HLA-I序列的蛋白组成氨基酸进行编码作为HLA-I序列特征;

[0031] 抗原受体-肽段互作特征通过以下方法获得标准特征:基于AAindex数据库中的AAindex3特征,进行序列比对,获取比对分数,通过尺度缩放获得标准化特征,作为抗原受体-肽段互作特征;

[0032] 肽段物理特征通过尺度缩放获得标准特征,以保证模型训练优化过程的数值稳定性;

[0033] 肽段氨基酸组分特征通过以下方法获得标准特征:统计肽段中标准氨基酸编码出现的次数,作为肽段氨基酸组分特征。

[0034] 具体的,抗原受体-肽段互作特征的尺度变化公式为:

[0035] $$p' = \frac{p - \min(p)}{\max(p) - \min(p)}$$

[0036] 式中,p表示抗原受体-肽段互作分数,p′表示作为标准特征的抗原受体-肽段互作分数。

[0037] 具体的,肽段物理特征的尺度变化公式为:

[0038] $$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

[0039] 式中,x表示肽段物理特征分数,x′表示作为标准特征的肽段物理特征分数。

[0040] 具体的,S3特征融合中,所述线性融合操作包括点乘、或加和、或组合成特征矩阵。

[0041] 具体的,S4中构建优化模型:

[0042] $$min\frac{\sum_{n=1}^{N}\left(y_n-f_W(x_n)\right)^2}{N}$$

[0043] 式中,f是含可学习参数的预测模型;W表示该模型中可学习参数,包括各融合特征获取时方案权重;x_n表示输入的特定数据,y_n是训练数据中免疫原性可信度数值;N表示样本总数。

[0044] 优选的,S5中求解优化模型:多次遍历所有训练数据,利用基于随机梯度优化方法的优化器进行优化,得到最优的模型参数,获得预测模型f_W。

[0045] 优选的,采用Beta分布,对IEDB数据库中具有实验验证的免疫原性结果进行编码,转换为回归拟合问题,基于此来提高训练模型可靠性;所述训练数据经过包括重抽样、剔除负样本的数据处理,避免使用的训练的数据正负样本量不平衡的问题。

[0046] 本发明还公开了一种基于多模态深度编码的新抗原免疫原性预测系统,它包括:

[0047] -数据收集模块,整理IEDB数据库中具有免疫原性验证结果的肽段及其MHC-I配体数据对;

[0048] -数据集构建,根据整理的IEDB数据库中具有免疫原性验证结果的肽段及其MHC-I配体数据对,构建正负样本。

[0049] -数据建模模块,构建预测模型并求解预测模型。

[0050] 优选的,所述数据建模模块包括:

[0051] -特征选择模块,选定与肽段免疫原性相关的特征,作为待融合特征;

[0052] -归一化处理模块,将待融合特征进行变换和尺度缩放,获得标准特征;

[0053] -特征融合模块,将多模态的标准特征输入深度神经网络进行融合,得到肽段免疫原性最终特征分数;

[0054] -预测模型构建模块,构建包含最终特征分数的预测模型和优化模型;

[0055] -预测模型求解模块,计算获得最优参数的预测模型。

[0056] 优选的,所述特征融合模块中,采用可变长的维度的输入设计,以便未来加入更多新模态特征的接口。

[0057] 优选的,采用Beta分布,对IEDB数据库中具有实验验证的免疫原性结果进行编码,转换为回归拟合问题,基于此来提高训练模型可靠性;所述训练数据经过包括重抽样、剔除负样本的数据处理,避免使用的训练的数据正负样本量不平衡的问题。

[0058] 更优的,它还包括:

[0059] -测试模块,整理包含在文献中但不包含在数据集构建模块出现的数据,以待优化得到最优模型后,验证测试模型的对于未曾见过的免疫原性数据对的预测效果。

[0060] 本发明的有益效果

[0061] 本发明提出了一种基于多模态深度编码的抗原免疫原性预测方法和系统,包括:1)基于抗体结合部位信息,包含更多HLA等位基因个数;2)可变长肽段编码方式,涵盖主要HLA结合肽段长度;3)考虑抗原受体谱系对肽段免疫原性的影响;4)包含肽段和MHC序列的物理和氨基酸组成特征;5)采用Beta分布,对IEDB数据库中具有实验验证的免疫原性结果进行编码,转换为回归拟合问题,基于此来提高训练模型可靠性;6)基于归一化处理,最终能得到格式、维度统一,便于融合的特征向量;7)基于可变长的维度的输入设计,以便未来

加入更多新模态特征的接口。未来加入的新特征只要是能被现有机器学习方法进行编码的，理论上没有任何限制，这也是我们模型“可拓展性”优点的体现。

附图说明

[0062] 图1多模态深度编码的抗原免疫原性预测方法计算流程图

[0063] 图2多模态深度编码的抗原免疫原性预测系统总结构图

[0064] 图3为10折交叉验证AUC评估结果图

[0065] 图4为10折交叉验证AUC-PR评估结果图

具体实施方式

[0066] 下面结合实施例对本发明作进一步说明，但本发明的保护范围不限于此：

[0067] 如图1所示，本发明提出的多模态深度编码的抗原免疫原性预测系统分为四部分，下面针对数据收集，数据集构建，模型建构与优化和模型测试进行详细阐述。

[0068] (a)数据收集

[0069] 该模块为根据IEDB数据库(参考文献:Vita R,Mahajan S,Overton JA,et al.The Immune Epitope Database(IEDB):2018update.Nucleic Acids Res.2019;47(D1):D339-D343.doi:10.1093/nar/gky1006)公开资源，整理与HLA-I免疫原性有关结果，具体包括：

[0070] I.选定T-cellassay中的linearpeptide；

[0071] II.选定Host Organism为Homo sapiens；

[0072] III.选定拥有完整HLA等位基因命名的HLA-I等位基因型；

[0073] IV.选取肽段长度为9或者10；

[0074] V.选定具有QualitativeMeasure信息的数据行；过滤缺失NumberofSubjectsTested或NumberofSubjectsResponded数值的数据行；过滤重复及结论冲突的数据行；

[0075] (b)数据集构建

[0076] 基于(a)中的方法收集HLA-I免疫原性的正负样本，构建模型训练数据集，具体如下：

[0077] I.使用抗体结合部位序列，代表HLA-I序列；

[0078] II.采用主成分分解PCA，对AAindex数据库中的AAindex1特征进行变换，选取变换后的12个主成分特征，对肽段和HLA中组分氨基酸进行编码；

[0079] III.使用开源计算R包Peptides(参考文献:Osorio D,Rondon-Villarreal P,Torres R(2015).“Peptides:A Package for Data Mining of Antimicrobial Peptides.”The R Journal,7(1),4-14.ISSN2073-4859.)，计算肽段的序列的电荷、疏水性、不稳定性物理特征，并进行尺度缩放；

[0080] IV.基于AAindex数据库中的AAindex3特征，将肽段与公开人类TCR数据集进行序列比对，获取比对分数；获取每一个肽段对应比对值分数的平均值，并进行尺度缩放；

[0081] V.统计每一个肽段对应的标准氨基酸编码个数。

[0082] VI.基于(a)中的方法收集HLA-I免疫原性的正负样本中，QualitativeMeasure取值为以下五种类型:Negative,Positive,Positive-High,Positive-Intermediate,

Positive-Low,分别对应:无免疫原性;有免疫原性;强免疫原性;中等免疫原性及弱免疫原性,基于其对应的Number of Subiects Tested和Number of Subjects Responded信息,使用如下Beta分布,生成10000个随机数,取这些值的均值作为免疫原性分数,进行免疫原性编码:

[0083] $$immu_{score}=\begin{cases}Beta\left(3+S,\ 3+(T-S)\right), & Negative\\ Beta(26+S,1+(T-S)), & Positive-Low\\ Beta\left(28+S,1+(T-S)\right), & Positive-Intermediate\\ Beta\left(30+S,1+(T-S)\right), & Positive\\ Beta\left(32+S,1+(T-S)\right), & Positive-High\end{cases}$$

[0084] 其中,T表示Number of Subjects Tested值;S表示Number of Subjects Responded值。以$immu_{score}$值表征每行数据免疫原性程度。

[0085] 选取以下文章中免疫原性结果,构建公共验证数据集,公共验证数据集数据集HLA及肽段编码方式同模型训练数据集:

[0086] 1)TESLA数据集:Wells DK,van Buuren MM,Dang KK,et al.Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction.Cell.2020;183(3):818-834.e13.doi:10.1016/j.cell.2020.09.015

[0087] 2)Emma数据集:Jappe EC,Garde C,Ramarathinam SH,et al.Thermostability profiling of MHC-bound peptides:a new dimension in immunopeptidomics and aid for immunotherapy design.Nat Commun.2020;11(1):6305.Published 2020 Dec 9.doi:10.1038/s41467-020-20166-4

[0088] 3)Ott数据集:Ott PA,Hu Z,Keskin DB,et al.An immunogenic personal neoantigen vaccine for patients with melanoma[published correction appears in Nature.2018 Mar14;555(7696):402].Nature.2017;547(7662):217-221.doi:10.1038/nature22991

[0089] 4)Bulik-Sullivan数据集:Bulik-Sullivan,B.,Busby,J.,Palmer,C.et al.Deep learningusing tumor HLA peptide mass spectrometry datasets improves neoantigen identification.Nat Biotechnol 37,55-63(2019).doi.org/10.1038/nbt.4313

[0090] 5)Robbins数据集:Robbins PF,Lu YC,El-Gamil M,et al.Mining exomic sequencing data to identifymutated antigens recognized by adoptively transferred tumor-reactive T cells.Nat Med.2013;19(6):747-752.doi:10.1038/nm.3161

[0091] (c)基于深度编码和多模态数据的模型建构与优化

[0092] 如图2的计算流程图所示,我们对模块(b)中的数据集进行编码、并建立模型。具体实施如下:

[0093] I.选择K-折交叉验证(k-fold cross validatio)的统计学方法构建模型训练、测试数据集。

[0094] II.构建深度学习模型,模型结构如图2。将肽段氨基酸组成、肽段物理特征、TCR-

肽段比对、编码肽段和编码HLA输入特征融合层，获得如下优化模型：

[0095] $$min\frac{\sum_{n=1}^{N}\left(y_n - f_W(x_n)\right)^2}{N}$$

[0096] 其中，f是含可学习参数的预测模型；W表示该模型中可学习参数，具体包括循环编码肽段；x_n表示输入的特定数据，y_n是训练数据中免疫原性可信度数值，既$immu_{score}$值；N表示样本总数。

[0097] III.最优模型的求解，采用批次随机梯度下降策略(参考文献:Goyal,Priya,et al."Accurate.large minibatch sgd:Training imagenet in 1 hour."arXiv preprint arXiv:1706.02677(2017).)：在多个轮次中，将训练数据分批次输入模型，计算如上的损失函数与梯度，并利用梯度下降更新模型。具体来说，我们采用ADMA优化器(参考文献：Kingma,Diederik P.,and Jimmy Ba."Adam:A method for stochastic optimization."arXiv preprint arXiv:1412.6980(2014).)，其用一阶梯度估计高阶梯度，并能自动调节优化的步长，是模型优化过程更加稳定与稳健。

[0098] (d)模型测试与机器学习评价指标

[0099] I.模型评估

[0100] 评估和交叉验证是测量模型性能的标准方法。它们都生成可检查或与其他模型比较的评估指标。

[0101] 我们采用接收者操作特征曲线(receiver operating characteristic curve，简称ROC曲线)下面积AUC与准确度-召回率(Precision-Recall，简称PR)曲线下面积AUC-PR来评价优化后模型的预测能力与性能(表1)：

[0102] 表1预测模型评价指标

[0103]

评价指标	描述
AUC	ROC曲线下而积
AUC-PR	PR曲线下面积

[0104] 此处，采用10折交叉验证，将模型训练数据集拆分为10部分，其中一个部分保留用于测试，其他9部分用于训练。此过程重复10次。

[0105] 如图3为10折交叉验证AUC评估结果，如图4为10折交叉验证AUC-PR评估结果。可以看到，模型在每一折上都具有很高的AUC值和AUC-PR值，均值分别为0.82和0.8。证明模型具有很好的泛化能力，能够很好的应对实际生产研发中的预测问题。

[0106] II.模型对比

[0107] 1.基于真实数据集TESLA数据集，将模型与两个广泛使用的免疫原性预测模型IEDB和DeepHLApan进行比较。评价指标设定如下(表2)：以0.5为阈值，计算灵敏度与精准度PPV，评估结果见表3。根据打分结果降序排列，我们分别选取了前20(Top20)及前50(Top50)计算PPV，可见本模型immu-D的结果是三种方法中最优的。基于全部数据，计算灵敏度，本模型immu-D也是最高的，两倍于DeepHLApan的结果。

[0108] 表2模型比较评估指标-1

[0109]

评价指标	描述
灵敏度	真阳性/(真阳性+假阴性)

精准度/PPV	真阳性/(真阳性+假阳性)

[0110] 表3模型比较评估结果-1

[0111]

	PPV		灵敏度
	Top20	Top50	
IEDB	1	3	0.63

[0112]

DeepHLApan	3	4	0.34
immu-D	4	6	0.74

[0113] 2.基于真实数据集Emma数据集，将模型与其文中提到的模型StabilityPredictor,MixMHCpred,NetMHCpan-4.0(EL),NetMHCpan-4.0(BA),MHCFlurry进行比较。评价指标设定如下(表4)：以0.5为阈值，计算AUC与精准度PPV，评估结果见表3。根据打分结果降序排列，我们分别选取了前10(Top10)计算PPV。结果表明(表5)，在AUC基本持平的情况下，本模型immu-D与基于复杂实验结果构建的模型StabilityPredictor具有相同的PPV值，远优于其余四个模型结果。

[0114] 表4模型比较评估指标-2

[0115]

评价指标	描述
AUC	ROC曲线下面积
精准度/PPV	真阳性/(真阳性+假阳性)

[0116] 表5模型比较评估结果-2

[0117]

	PPV-Top10	AUC
Stability Predictor	0.9	0.75
MixMHCpred	0.7	0.7
NetMHCpan-4.0(EL)	0.6	0.68
NetMHCpan-4.0(BA)	0.6	0.67
MHCFlurry	0.7	0.65
immu-D	0.9	0.66

[0118] III.真实数据结果展示

[0119] 真实生产环境下，由于相关的限制，通常只会对部分候选肽段进行后续实验验证。为了验证模型在实际生产中的作用，我们以精准度PPV为判别指标，分别计算模型在前10(Top10)、前20(Top20)、前30(Top30)及整体数据上的表现，结果见表6。从结果可知，即使在高阳性-阴性比数据上，我们的模型也能很好的捕获阳性结果。其也佐证了我们模型在真实生产环境下的价值。

[0120] 表6真实数据结果

[0121]

				PPV			
数据集	样本数	阳性数	阴性数	Top10	Top20	Top30	overall
Ott	144	15	129	0.2	0.2	0.13	0.11
Bulik-Sullivan	72	8	64	0.3	0.15	0.13	0.11
Robbins	217	9	208	0.1	0.5	0.8	0.7

[0122] 应当理解的是，本发明的应用不限于上述的据力。对本领域从业技术人员来说，可以根据上述说明加以改进或者变换，特别是基本模型选取、免疫指标构建方法及相关特征值的添加。所有这些改进和变换，以及参数相关的调节和选取，都应属于本发明所附权利要求的保护范围。

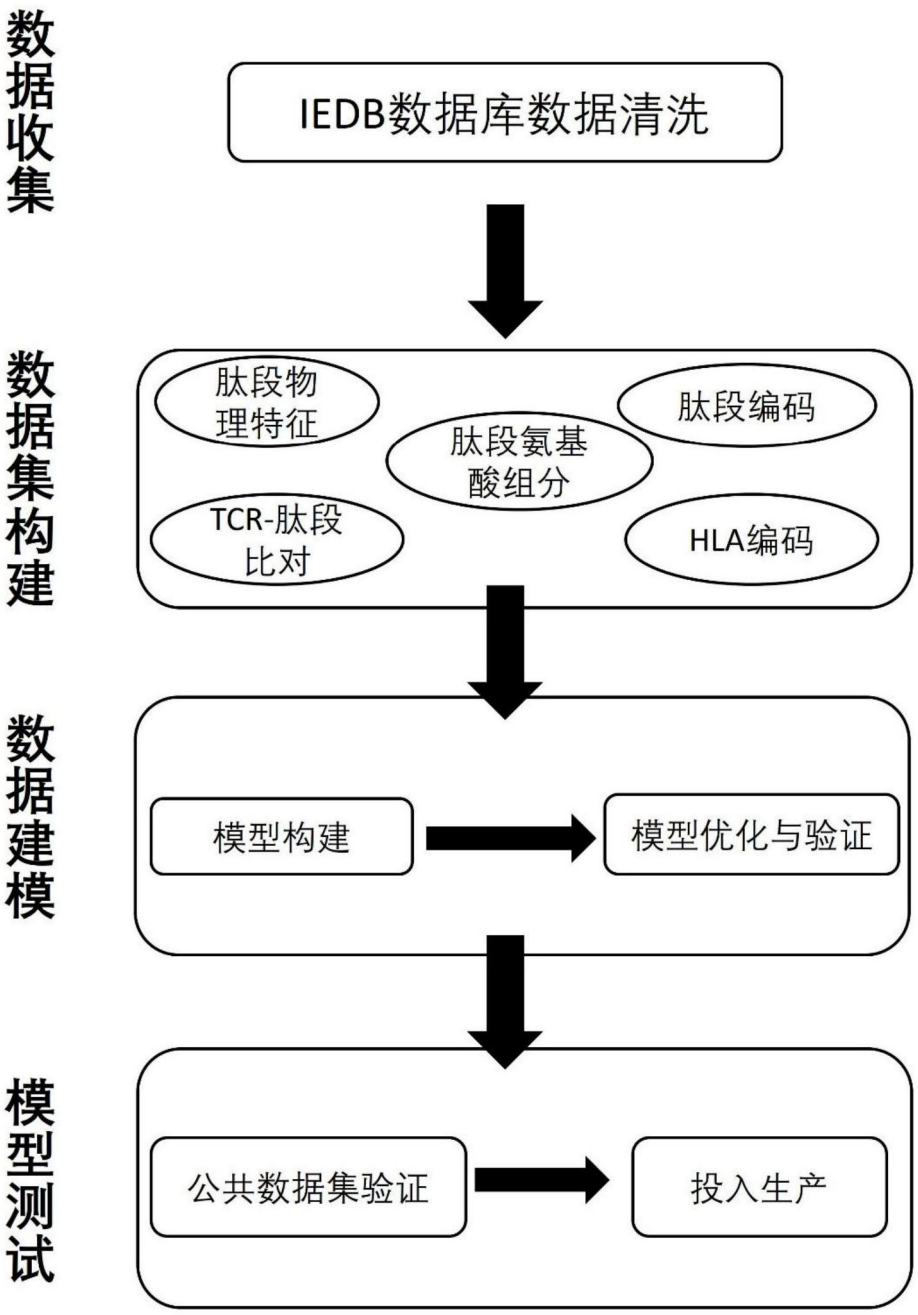

数据收集
IEDB数据库数据清洗
数据集构建
肽段物理特征
肽段氨基酸组分
肽段编码
TCR-肽段比对
HLA编码
数据建模
模型构建
模型优化与验证
模型测试
公共数据集验证
投入生产

图1

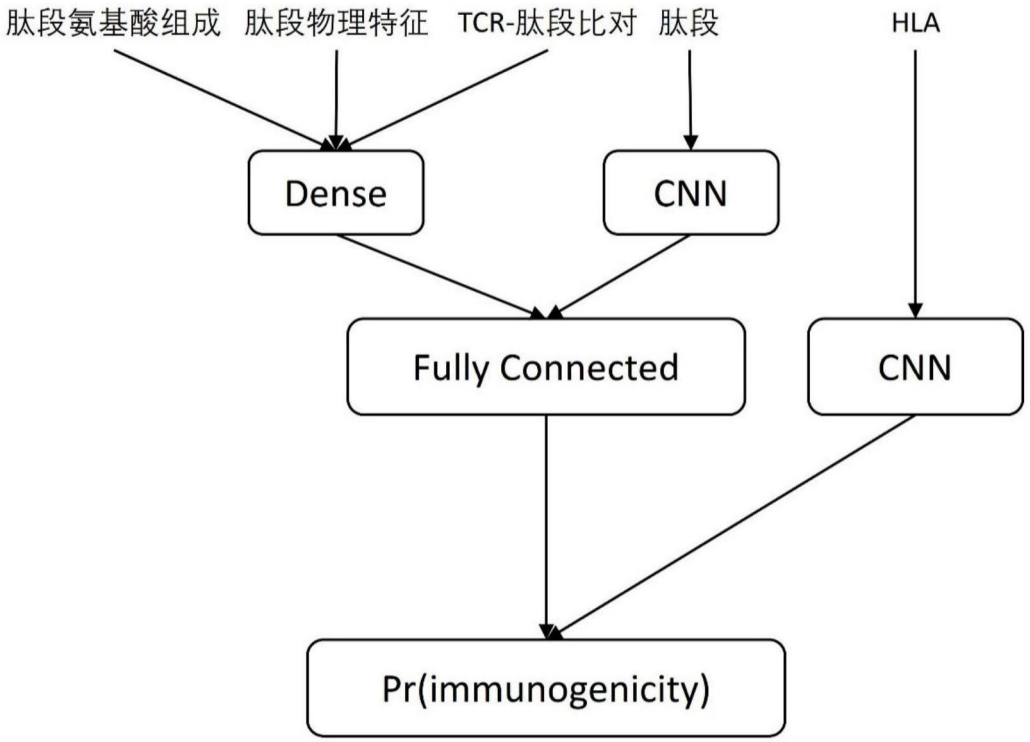

肽段氨基酸组成
肽段物理特征
TCR-肽段比对
肽段
HLA
Dense
CNN
Fully Connected
CNN
Pr(immunogenicity)

图2

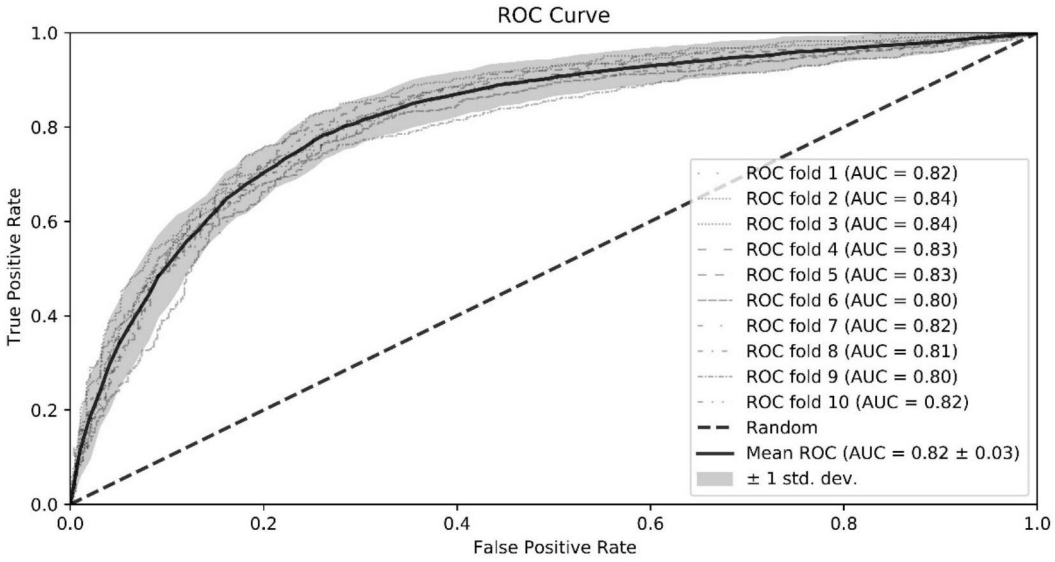

ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 1 (AUC = 0.82)
ROC fold 2 (AUC = 0.84)
ROC fold 3 (AUC = 0.84)
ROC fold 4 (AUC = 0.83)
ROC fold 5 (AUC = 0.83)
ROC fold 6 (AUC = 0.80)
ROC fold 7 (AUC = 0.82)
ROC fold 8 (AUC = 0.81)
ROC fold 9 (AUC = 0.80)
ROC fold 10 (AUC = 0.82)
Random
Mean ROC (AUC = 0.82 ± 0.03)
± 1 std. dev.

图3

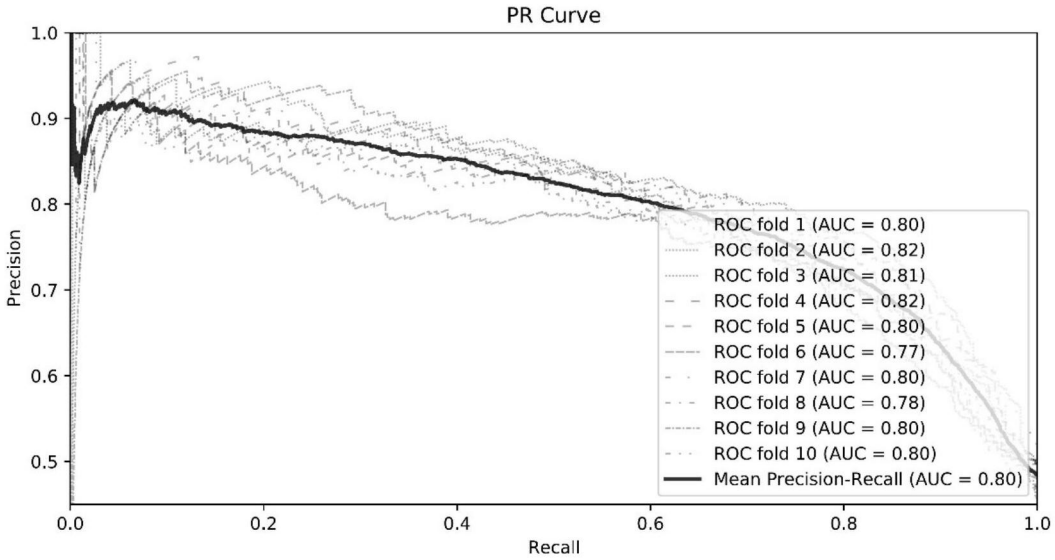

PR Curve
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 1 (AUC = 0.80)
ROC fold 2 (AUC = 0.82)
ROC fold 3 (AUC = 0.81)
ROC fold 4 (AUC = 0.82)
ROC fold 5 (AUC = 0.80)
ROC fold 6 (AUC = 0.77)
ROC fold 7 (AUC = 0.80)
ROC fold 8 (AUC = 0.78)
ROC fold 9 (AUC = 0.80)
ROC fold 10 (AUC = 0.80)
Mean Precision-Recall (AUC = 0.80)

图4