

一种用于鉴定与 MHC-I/HLA-I 类结合及 TCR 识别肽段的方法、设备及存储介质

技术领域

本发明涉及疫苗领域，特别涉及一种用于鉴定与 MHC-I/HLA-I 类结合及 TCR 识别肽段的方法、设备及存储介质。

背景技术

抗原筛选技术主要是筛选出在疫苗制备过程中能够产生最强的免疫反应的抗原。在抗原筛选过程中，需要对不同来源的抗原进行筛选，并确定最佳抗原组合，常用的抗原筛选技术包括基于生物芯片的筛选技术和基于蛋白质分析的筛选技术等。肿瘤疫苗是目前疫苗领域中的研究热点，最早的肿瘤疫苗见于 1994 年，近年来技术的进步使肿瘤疫苗可以靶向治疗患者特异性突变抗原，通过将肿瘤相关抗原或者肿瘤特异性抗原的肽段或 mRNA 注入患者体内，来激活患者的免疫系统。个性化肿瘤疫苗技术的突破，关键在于新生抗原的检测及预测技术的发展，以制备新生抗原疫苗。预测特异性突变肽与组织相容性复合体的结合潜力是新生抗原检测的一个重要组成部分，因为肿瘤特异性抗原只有被主要组织相容性复合物识别呈递至细胞表面，才能进一步被 T 细胞受体识别，并最终杀死肿瘤细胞。因此，预测肽段和 MHC-I/HLA-I 的结合，以及结合后的复合物 pMHC 能够保持稳定的结合状态，能否被呈递至细胞表面并被 TCR 识别具有重要意义。

近年来，新生抗原领域的发展非常活跃，产生广泛和大量的数据库可供使用，用于进行肽段与 MHC-I/HLA-I 亲和力预测模型的训练。同时多种算法被应用于肽段与 MHC-I/HLA-I 结合潜力预测，包括传统的机器学习方法及发展势头迅猛的深度学习方法，人工神经网络 ANN、卷积神经网络算法 CNN、贝叶斯最小二乘法 SMM，随机森林等方法都曾被用于亲和力模型构建。虽然这些算法在测试集测试时都被报道具有良好的新能，但是实际更换完全独立的数据集时却表现较为一般，肽段和 MHC-I/HLA-I 的亲和力预测仍然有很高的错误率，特别是假阳性率，这将在临床应用时产生严重的影响。产生此问题的原因在于以下几点：一是无法获得全面且大量的用于模型训练的数据集，从有限的数据集中很难学到更多的特征；二是应用于预测模型的常规算法较少的考虑到蛋白序列本身的特征，及可能在蛋白序列互作时产生重要影响的序列特征；三是无论是 MHC-I/HLA-I 全长序列的表示方式还是 MHC-I/HLA-I 伪序列的表征方式都使得模型无法有针对性的学习肽段结合中重要位点的特征，影响模型性能；四是目前常见的方法都只关注到 pMHC 呈递，突变肽段的特征例如结合位置，氨基酸残基类型，并没有进一步考虑 pMHC 结合稳定性、pMHC 呈递能力和中枢耐受机制等在 TCR 识别时产生重要

影响，影响进一步选择更为准确的新生抗原突变肽。

发明内容

为解决上述问题，本发明提供一种用于鉴定与 MHC-I/HLA-I 类结合及 TCR 识别肽段的方法、设备及存储介质，更有针对性的获取肿瘤特异性肽段的识别，转运、呈递及 TCR 识别的关键特征的方法，并应对模型构建数据量不足的现状，应用新的建模思想进行模型训练和特征提取及分组，实现鉴定与 MHC-I/HLA-I 类结合的肽段及新生抗原准确性的提升。

本发明实施例提供了一种用于鉴定与 MHC-I/HLA-I 类结合及 TCR 识别肽段的方法，包括：

获取蛋白质序列、MHC-I/HLA-I 序列和肽段序列；

利用所述蛋白质序列训练蛋白质序列分词器；

利用经过所述蛋白质序列分词器进行分词处理后的蛋白质序列进行蛋白质表征模型训练，得到训练后的蛋白质表征预训练模型；

利用经过所述蛋白质序列分词器进行分词处理后的 MHC-I/HLA-I 序列和肽段序列进行拼接处理，对所述训练后的蛋白质表征预训练模型进行微调，得到二分类模型；

利用所述二分类模型对目标野生型/突变型肽段进行鉴定处理，对突变型肽段进行分类，获取与 TCR 识别正相关的候选新生抗原突变肽；

二分类模型可应用于任何可与 MHC-I/HLA-I 结合的肽段的预测，候选新生抗原可通过应用二分类模型和突变所发生的位置来预测。

本发明实施例提供了一种用于鉴定与 MHC-I/HLA-I 类结合及 TCR 识别肽段的方法，包括：

获取蛋白质序列、MHC-I/HLA-I 序列和肽段序列；

利用所述蛋白质序列训练蛋白质序列分词器；

利用经过所述蛋白质序列分词器进行分词处理后的蛋白质序列进行蛋白质表征模型训练，得到训练后的蛋白质表征预训练模型；

利用经过所述蛋白质序列分词器进行分词处理后的 MHC-I/HLA-I 序列和肽段序列进行拼接处理，对所述训练后的蛋白质表征预训练模型进行微调，得到二分类模型；

利用所述二分类模型对待鉴定的肽段进行鉴定，获取与 MHC-I/HLA-I 结合的肽段。

其中，所述 MHC-I 是指主要组织相容性复合物 I 类；所述 HLA-I 是指人类中的主要组织相容性复合物 I 类；所述 TCR 是指 T 细胞受体。

优选地，所述肽段序列包括阳性肽段序列和阴性肽段序列，其中，所述阳性肽段序列为与 MHC-I/HLA-I 结合的肽段，标签为 1；所述阴性肽段序列为根据阳性肽段序列特征在蛋白序列中随机生成的肽段序列，标签为 0。

优选地，所述蛋白质序列包括从蛋白质数据库中收集蛋白质序列；相应地，利用所收集的蛋白质序列进行高频重复氨基酸残基组合 motif 的分词器训练，获得蛋白质序列分词器。

优选地，所述利用经过所述蛋白质序列分词器进行分词处理后的蛋白质序列进行蛋白质表征模型训练，得到训练后的蛋白质表征预训练模型，包括：

将无标签的蛋白质序列输入至蛋白质序列分词器中进行分词处理，得到氨基酸功能残基组合 motif 分割后的蛋白质序列；

将所述氨基酸功能残基组合 motif 分割后的蛋白质序列转换为数字表示的蛋白质序列；

利用所述数字表示的蛋白质序列进行蛋白质表征模型训练，得到训练后的蛋白质表征预训练模型。

优选地，所述利用经过所述蛋白质序列分词器进行分词处理后的 MHC-I/HLA-I 序列和肽段序列对所述训练后的蛋白质表征预训练模型进行微调，得到二分类模型包括：

将带标签的 MHC-I/HLA-I 序列输入至蛋白质序列分词器中进行分词处理，得到氨基酸功能残基组合 motif 分割后的 MHC-I/HLA-I 序列，并将所述氨基酸功能残基组合 motif 分割后的 MHC-I/HLA-I 序列转换为数字表示的 MHC-I/HLA-I 序列；

将带标签的肽段序列转换为数字表示的肽段序列；

通过将转换得到拼接后的序列，并利用所述拼接后的序列对所述训练后的蛋白质表征预训练模型进行微调，得到二分类模型。

在具体的实施方式中，微调可包括继承原本的预训练模型并增加分类层；此外，可以在增加分类层的基础上进一步增加全连接层、注意力机制层或其他合适的任务层等。

在一些实施方式中，拼接策略为直接拼接，即直接将数字表示的 MHC-I/HLA-I 序列和数字表示的肽段序列进行拼接。

在另一些实施方式中，拼接策略为将数字表示的 MHC-I/HLA-I 序列、数字表示的肽段序列进行拼接，以及数字表示的与肽段和 MHC-I/HLA-I 结合相关的其他特征进行拼接。本领域技术人员可以根据拼接对象选择合适的拼接方式，示例性的拼接方式是直接将数字表示的序列和相关特征 `torch.cat()` 或者按照自定义的拼接方式进行特征张量的加减乘除等。

在具体的实施方式中，与肽段和 MHC-I/HLA-I 结合相关的其他特征选自蛋白质序列的裂解可能性、MHC-I/HLA-I 与肽段的结合稳定性，和 MHC-I/HLA-I 与肽段复合物呈递可能性中的一种或多种。

蛋白质序列的裂解可能性可以使用合适的软件，输入肽段序列后得到。本领域技术人员可以理解，可以使用任何合适的软件及其合适的版本，只要能够根据肽段序列得到蛋白质序列的裂解可能性即可，例如 `pepsickle` 和 `iPCPS` 软件等可以达到此目的。

MHC-I/HLA-I 与肽段的结合稳定性可以使用合适的软件, 输入肽段序列及 MHC-I/HLA-I 序列后得到。本领域技术人员可以理解, 可以使用任何合适的软件及其合适的版本, 只要能够根据肽段序列及 MHC-I/HLA-I 序列得到 MHC-I/HLA-I 与肽段的结合稳定性即可, 例如 LKMHCstablepan 和 TLStab 软件等可以达到此目的。

MHC-I/HLA-I 与肽段复合物呈递可能性可以使用合适的软件, 输入肽段序列后得到。本领域技术人员可以理解, 可以使用任何合适的软件及其合适的版本, 只要能够根据肽段序列及 MHC-I/HLA-I 序列(可选)得到 MHC-I/HLA-I 与肽段复合物呈递可能性即可, 例如 DeepTAP、TAPPred、TAPREG 软件等可以达到此目的。

优选地, 在微调的步骤之前, 本发明的方法进一步包括利用蛋白质序列、MHC-I/HLA-I 序列和肽段序列计算与肽段和 MHC-I/HLA-I 结合相关的其他特征。

优选地, 所述利用所述二分类模型对目标野生型/突变型肽段进行鉴定处理包括:

将目标野生型/突变型肽段输入至所述二分类模型中, 输出预测为阳性的突变型肽段每个位点的注意力得分, 以及 MHC-I/HLA-I 序列的氨基酸功能残基组合 motif 的注意力得分;

根据所述阳性突变型肽段每个位点的注意力得分以及 MHC-I/HLA-I 序列 motif 的注意力得分, 判断在 MHC-I/HLA-I 序列与阳性肽段结合时的锚定位以及 MHC-I/HLA-I 序列的关键表征位点, 并利用所述 MHC-I/HLA-I 序列与肽段结合时的锚定位点和肽段的突变位点对阳性突变型肽段进行分类处理。

优选地, 所述利用所述 MHC-I/HLA-I 序列与阳性突变型肽段结合时的锚定位以及相对于野生型肽段的突变位进行分类处理包括:

根据所述 MHC-I/HLA-I 序列与肽段结合时的锚定位点, 并结合突变所在位置进行阳性突变型肽段分类处理, 并根据分类结果判断是否适合被选为 TCR 识别正相关的候选新生抗原突变肽。

优选地, 所述根据分类结果判断是否适合被选为 TCR 识别正相关的候选新生抗原突变肽包括:

当野生型肽段为弱 MHC-I/HLA-I 结合且弱 TCR 识别(一般情况下野生型肽段均为弱 TCR 识别), 突变型肽段为 MHC-I/HLA-I 锚定位点突变, 转为强 MHC-I/HLA-I 和强 TCR 识别时, 则判断能被选为新生抗原突变肽;

当野生型肽段为弱 MHC-I/HLA-I 结合且弱 TCR 识别(一般情况下野生型肽段均为弱 TCR 识别), 突变型肽段为非 MHC-I/HLA-I 锚定位点突变, 转为强 MHC-I/HLA-I 和强 TCR 识别时, 则判断能被选为新生抗原突变肽;

当野生型肽段为强 MHC-I/HLA-I 结合且弱 TCR 识别(一般情况下野生型肽段均为弱 TCR 识别), 突变型肽段为非 MHC-I/HLA-I 锚定位点突变, 转变为强 MHC-I/HLA-I 和强 TCR 识别(这里的强 TCR 识别指暴露给 TCR 的序列发生改变)时, 则判断能被

选为新生抗原突变肽；

当野生型肽段为强 MHC-I/HLA-I 结合且弱 TCR 识别（一般情况下野生型肽段均为弱 TCR 识别），突变型肽段为 MHC-I/HLA-I 锚定位点突变，转为强 MHC-I/HLA-I 和弱 TCR 识别（这里的弱 TCR 识别指暴露给 TCR 的序列未发生改变）时，则判断不能被选为新生抗原突变肽。

本发明的有益效果是，通过应用自然语言处理的深度学习算法，鉴定野生型/突变型肽段与组织相容性分子复合物 MHC-I/HLA-I 的识别，转运及呈递，识别蛋白质相互作用时的生物学特征，判断肽段与 MHC-I/HLA-I 的锚定位点以及构建 MHC-I/HLA-I 更具有代表性的序列表征方式，同时引入中枢耐受机制对 TCR 识别的影响，进一步推动肿瘤新生抗原的识别。

附图说明

图 1 是本发明提供的一种用于鉴定与 MHC-I/HLA-I 类结合及 TCR 识别的肽段的鉴定方法的流程图；

图 2 是本发明提供的一种用于鉴定与 MHC-I/HLA-I 类结合及 TCR 识别的肽段的算法的流程图；

图 3 是本发明提供的一种具有其他特征的用于鉴定与 MHC-I/HLA-I 类结合及 TCR 识别的肽段的算法的流程图。

具体实施方式

应当理解，此处所描述的具体实施例仅用以解释本发明，并不用于限定本发明。在后续的描述中，使用用于表示元件的诸如“模块”、“部件”或“单元”的后缀仅为了有利于本发明的说明，其本身没有特有的意义。因此，“模块”、“部件”或“单元”可以混合地使用。

如本文所使用的，术语“拼接处理”是指将转换为数字表示的多个特征进行拼接，从而得到能对蛋白质表征预训练模型进行微调的拼接后序列的过程。

如本文所使用的，术语“pMHC”是指肽段和 MHC-I/HLA-I 结合后的复合物。

如本文所使用的，术语“其他特征”与“更多的特征”可互换使用，指利用蛋白质序列、MHC-I/HLA-I 序列和肽段序列计算的与肽段和 MHC-I/HLA-I 结合相关的特征。示例性的特征是蛋白质序列的裂解可能性、pMHC 结合稳定性和 pMHC 呈递能力。

本发明提供了用于鉴定野生型/突变型肽段是否与主要组织相容性复合物 I 类 MHC-I/HLA-I 结合、转运、呈递至细胞表面及 TCR 识别的可能性的深度学习算法，进一步推动鉴定与 MHC-I/HLA-I 类结合的肽段和肿瘤特异性新生抗原的选择。具体的，本方法为：将自然语言处理的方法应用于蛋白质序列特征发现和提取，用于构建蛋白质表

征预训练模型；利用蛋白序列及蛋白互作数据集，提取与肽段结合，转运和呈递正相关的氨基酸残基组合 motif 及关键位点特征，在蛋白质表征预训练模型的基础上，通过收集肽序列阳性集及构建阴性集进行蛋白质表征预训练模型微调，获得准确的可推断野生型/突变型肽段与 MHC-I/HLA-I 结合可能性的模型。上述方法获得的模型可以应用于泛类型的肽段或 MHC-I/HLA-I 类型结合预测，启发性的反映肽段呈递的生物学特征，并将其应用于模型开发，获得比常规算法更具解释性和准确性的预测模型。同时进一步的识别肽段与 MHC-I/HLA-I 结合之后的 pMHC 与 TCR 的识别特征，从而进一步预测肽段是否具有激活 TCR 的能力，而不仅仅是与 MHC 结合，最终可用于选择更具潜力的候选新生抗原。

图 1 是本发明提供的一种用于鉴定与 MHC-I 类结合及 TCR 识别的肽段的方法的流程图，如图 1 所示，包括：

步骤 S101：获取蛋白质序列、MHC-I/HLA-I 序列和肽段序列；

步骤 S102：利用所述蛋白质序列训练蛋白质序列分词器；

步骤 S103：利用经过所述蛋白质序列分词器进行分词处理后的蛋白质序列进行蛋白质表征模型训练，得到训练后的蛋白质表征预训练模型；

步骤 S104：利用经过所述蛋白质序列分词器进行分词处理后的 MHC-I/HLA-I 序列和肽段序列对所述训练后的蛋白质表征预训练模型进行微调，得到二分类模型；

步骤 S105：利用所述二分类模型对目标野生型/突变型肽段进行鉴定处理，对突变型肽段进行分类，获取与 TCR 识别正相关的候选新生抗原突变肽。

具体地说，所述肽段序列包括阳性肽段序列和阴性肽段序列，其中，所述阳性肽段序列为与 MHC-I/HLA-I 结合的肽段，标签为 1；所述阴性肽段序列为根据阳性肽段序列特征在蛋白序列中随机生成的肽段序列，标签为 0。

其中，所述蛋白质序列包括从蛋白质数据库中收集蛋白质序列；相应地，利用所收集的蛋白质序列进行高频重复氨基酸残基组合 motif 的分词器训练，获得蛋白质序列分词器。

进一步地，所述利用经过所述蛋白质序列分词器进行分词处理后的蛋白质序列进行蛋白质表征模型训练，得到训练后的蛋白质表征预训练模型，包括：将无标签的蛋白质序列输入至蛋白质序列分词器中进行分词处理，得到氨基酸功能残基组合 motif 分割后的蛋白质序列；将所述氨基酸功能残基组合 motif 分割后的蛋白质序列转换为数字表示的蛋白质序列；利用所述数字表示的蛋白质序列进行蛋白质表征模型训练，得到训练后的蛋白质表征预训练模型。

更具体地说，所述利用经过所述蛋白质序列分词器进行分词处理后的 MHC-I/HLA-I 序列和肽段序列对所述训练后的蛋白质表征预训练模型进行微调，得到二分类模型包括：将带标签的 MHC-I/HLA-I 序列输入至蛋白质序列分词器中进行分词处理，得到氨基酸

功能残基组合 motif 分割后的 MHC-I/HLA-I 序列，并将所述氨基酸功能残基组合 motif 分割后的 MHC-I/HLA-I 序列转换为数字表示的 MHC-I/HLA-I 序列；将带标签的肽段序列转换为数字表示的肽段序列；通过将所述转换为数字表示的 MHC-I/HLA-I 序列和所述转换为数字表示的肽段序列进行拼接处理，得到拼接后的序列，拼接策略包括：直接拼接，或者加入其他数字表示的特征，包括蛋白质序列的裂解可能性、MHC-I/HLA-I 与肽段的结合稳定性，和 MHC-I/HLA-I 与肽段复合物呈递可能性中的一种或多种，并利用所述拼接后的序列对所述训练后的蛋白质表征预训练模型进行微调，得到二分类模型。

进一步地，所述利用所述二分类模型对目标野生型/突变型肽段进行鉴定处理包括：将目标野生型/突变型肽段输入至所述二分类模型中，输出预测为阳性的突变型肽段每个位点的注意力得分，以及 MHC-I/HLA-I 序列的氨基酸功能残基组合 motif 的注意力得分；根据所述阳性突变型肽段每个位点的注意力得分以及 MHC-I/HLA-I 序列 motif 的注意力得分，判断在 MHC-I/HLA-I 序列与阳性肽段结合时的锚定位以及 MHC-I/HLA-I 序列的关键表征位点，并利用所述 MHC-I/HLA-I 序列与肽段结合时的锚定位点和肽段的突变位点对阳性突变型肽段进行分类处理。

进一步地，所述利用所述 MHC-I/HLA-I 序列与阳性突变型肽段结合时的锚定位以及相对于野生型肽段的突变位进行分类处理包括：根据所述 MHC-I/HLA-I 序列与肽段结合时的锚定位点，并结合突变所在位置进行阳性突变型肽段分类处理，并根据分类结果判断是否适合被选为 TCR 识别正相关的候选新生抗原突变肽。

更具体地说，所述根据分类结果判断是否适合被选为 TCR 识别正相关的候选新生抗原突变肽包括：当野生型肽段为弱 MHC-I/HLA-I 结合且弱 TCR 识别，突变型肽段为 MHC-I/HLA-I 锚定位点突变，转为强 MHC-I/HLA-I 和强 TCR 识别时，则判断能被选为新生抗原突变肽；当野生型肽段为弱 MHC-I/HLA-I 结合且弱 TCR 识别，突变型肽段为非 MHC-I/HLA-I 锚定位点突变，转为强 MHC-I/HLA-I 和强 TCR 识别时，则判断能被选为新生抗原突变肽；当野生型肽段为强 MHC-I/HLA-I 结合且弱 TCR 识别，突变型肽段为非 MHC-I/HLA-I 锚定位点突变，转变为强 MHC-I/HLA-I 和强 TCR 识别时，则判断能被选为新生抗原突变肽；当野生型肽段为强 MHC-I/HLA-I 结合且弱 TCR 识别，突变型肽段为 MHC-I/HLA-I 锚定位点突变，转为强 MHC-I/HLA-I 和弱 TCR 识别时，则判断不能被选为新生抗原突变肽。

本申请实施例提供的一种电子设备，包括：存储器；处理器；以及计算机程序；其中，所述计算机程序存储在所述存储器中，并被配置为由所述处理器执行以实现一种用于鉴定与 MHC-I/HLA-I 类结合及 TCR 识别的肽段的方法、设备及存储介质。

本申请实施例提供的一种计算机可读存储介质，其上存储有计算机程序；所述计算机程序被处理器执行以实现一种用于鉴定与 MHC-I/HLA-I 类结合及 TCR 识别的肽段的

方法。

综上所述，本发明提供了用于推断肽段与 MHC-I/HLA-I 结合可能性的方法，从蛋白质序列中学习生物特征，以预训练模型为基础，应用于获得具有解释性和准确性的肽段和 MHC-I/HLA-I 结合预测模型，具体包括：应用大量无标签的蛋白质数据构建蛋白质序列分词器和蛋白质表征预训练模型；构建包括阳性/阴性数据集的有标签的训练集，验证集和测试集；其中阳性数据集包括从数据库中获取的通过质谱或者体外结合实验获得的阳性肽序列和其对应的 MHC-I/HLA-I 亚型；其中阴性序列为根据阳性肽段序列长度特征在蛋白序列中随机生成的阴性肽序列，阴性肽段需剔除掉在阳性肽段中存在的肽段。对 MHC-I/HLA-I 亚型使用蛋白质序列分词器进行分词，并将 MHC-I/HLA-I 序列和肽段序列转换为数字表示；构建二分类模型，在蛋白质表征预训练模型的基础上进行微调，对生成的数字表示的数据进行有监督学习，进一步提高模型在该任务上的性能；进一步的应用二分类模型，根据其预测的阳性肽段与 MHC-I/HLA-I 序列 motif 的注意力得分，获得表征 MHC-I/HLA-I 方式，以及推断肽段与 MHC-I/HLA-I 的结合锚定位点；更深层次的，推出一种可推断肽段能否与 MHC-I/HLA-I 结合并激活 TCR 反应，激活免疫反应的方法，利用肽段与 MHC-I/HLA-I 结合的锚定位置，根据突变所在肽段的位置以及中枢耐受的影响，对野生型/突变型肽段与 MHC-I/HLA-I 结合以及 TCR 识别类型进行分类，根据分类结果对新生抗原肽段进行进一步的剔除，获取更为准确的新生抗原。

本发明涉及一种全新的预测野生型/突变型肽段与 MHC-I/HLA-I 结合，转运，呈递及激活 TCR 的方法，通过提取在结合过程中发挥重要作用的蛋白质残基组合和关键位点、更好的表征 MHC-I/HLA-I 序列以及应用迁移学习技术构建更为准确且具有生物学上可解释性的模型预测器。利用构建好的模型探索 MHC-I/HLA-I 关键结合位点并进一步的根据突变所在肽段的位置进行新抗原分类，提升新抗原识别的准确性。本方法与以往的分类器相比更为高效精准，对于肿瘤免疫疗法具有重要的应用价值。

如下文实施例中所证实的，相比其他已有的分析方法，具有如下优点：

(1) 将蛋白质序列本身的生物学特征应用于模型构建中。通过学习蛋白质序列中氨基酸残基组合方式，发现高频的具有显著特征的氨基酸残基组合和关键位点，构建分词器，更合理的表征氨基酸序列。

(2) 将自然语言处理算法应用于生物信息语言-蛋白序列中，将蛋白序列看作是一段语句，使用自然语言处理的方法构建预测模型，学习序列中的特征及上下文关系。

(3) 通过已有的大量无标签数据构建预训练模型，使用有标签数据进行预训练模型微调。通过构建掩码模型来推断被掩盖的氨基酸残基组合以学习蛋白质序列的上下文信息和表述规则，帮助预训练模型学习到丰富的语言知识。然后使用有限的有标签的肽段和 MHC-I/HLA-I 结合数据进行预训练模型微调，构建更适合亲和力预测的分类模型。

(4) 借助自然语言处理模型可输出蛋白质序列中每个位点/motif 的注意力得分，

以此判断在 MHC-I/HLA-I 与肽段结合时的关键位点，判断肽段的关键结合位点以及 MHC-I/HLA-I 序列的更具代表性和显著性的表征方式。

(5) 不仅预测肽段与 MHC-I/HLA-I 结合，也通过拼接涵盖肽段与 MHC-I/HLA-I 的结合稳定性，转运及呈递等特征，从实际生物学进程出发，将在生物学通路中起重要作用的特征转变为数字形式，并与其他特征相结合应用到模型训练和预测中，提高模型的实际应用能力。

(6) 通过判断肿瘤特异性突变所在肽段的位置，对不同的 pMHC 结合及 TCR 识别方式进行分组，进一步提取更具准确性的生抗原肽段。根据文献报道，抗原序列一面与 MHC 结合，一面被 TCR 识别，肿瘤特异性抗原突变所在的位置会同时影响 MHC 的结合和 TCR 的识别。因此，我们根据每个抗原肽位置的注意力得分获取关键的 MHC 结合位点，并以此进一步根据突变位置进行新抗原分类，剔除掉肽段与 MHC-I/HLA-I 结合潜力大但是却不能激活 TCR，提高抗原肽筛选的准确性。改进后的算法能更加精准的预测野生型/突变型肽段是否与主要组织相容性复合物的结合，转运及呈递和 TCR 识别潜力，推动肿瘤新抗原的鉴别及其临床应用。

下面通过结合附图 2 和图 3 对本次申请的内容进行详细说明

实施例一

1.1、模型训练所需的数据集并对其进行筛选

1.1.1、从 UniProt 等数据库中获取蛋白序列；

1.1.2、从 IEDB 等数据库中获取阳性肽序列，阳性肽段需同时符合以下条件：

肽段长度为 8-15bp 的阳性肽；

识别阳性肽的 MHC/HLA 类型为 I 型；

对应的野生型肽段有具体的 UniProt 编号。

1.1.3、从 IPD-IMGT/HLA 等数据库中获取 HLA 序列，并获取多序列比对后的 HLA-I 序列，用于表征 HLA 序列。

1.2、构建阳性数据集和阴性数据集

其中阳性数据集为蛋白质序列/阳性肽序列，其中蛋白序列用于构建预训练模型，学习蛋白序列的表达特征和上下文联系。阳性肽序列为带有标签的可以与 MHC-I/HLA-I 结合的肽段，阴性序列为在阳性序列的基础上随机构建而成，即根据阳性肽段推断出未与 MHC-I/HLA-I 结合的肽段，阴性肽段构建方法为：

肽段长度为 8-15bp 的阴性肽；

在阳性肽同一条蛋白序列基础上随机变换氨基酸产生；

剔除掉在阳性肽中出现的肽段。

1.3、构建并应用蛋白质序列分词器

1.3.1、通过机器学习算法或者统计推断模型对大量的蛋白质进行氨基酸残基组合，并

推断在蛋白质序列中高频出现的氨基酸残基组合,这些经常以组合形式出现的氨基酸残基很可能是功能 motif,或者说在蛋白质行使功能时发挥重要作用。同时,可推断重要 motif 的点位,区别于简单关注单个氨基酸残基,分词器更有利于帮助探索在蛋白质发挥功能时的关键特征。

1.3.2、蛋白质序列分词。所有蛋白质序列都经过分词器进行分词,将可能联合起作用的氨基酸残基组合到一起,以便于后续进行模型构建。

本发明实施例能在蛋白质数据库中学习蛋白质序列中氨基酸残基的组合方式,将蛋白质序列中高频组合残基组合为一组,构建蛋白质序列分词器,包括:从蛋白数据库中下载蛋白质数据集合,从中提取蛋白质序列;对蛋白质序列进行训练,根据已知蛋白质序列中氨基酸残基/残基组合的出现频率,左右连接度等特征,训练具有可解释性的蛋白质序列氨基酸残基分词器。

1.4、蛋白质表征预训练模型构建

大量的经过分词的蛋白质序列,隐藏着蛋白质功能相关特征,蛋白质表达方式,以及序列中 motif 之间的关系,因此使用自然语言处理中的掩码语言模型,随机掩盖蛋白质序列中的 motif,对其进行预测,学习蛋白质序列中所隐含的特征。掩码语言模型是一种自监督学习任务,旨在让模型在预训练时学习到 motif 的上下文信息和语言规则,模型接收输入的序列,然后在输入的文本序列中随机选择一定比例的 motif 进行掩盖,并要求模型预测这些被掩盖的 motif。例如,给定下面的句子:

输入序列: MAF SAE DVL KE YD RRRR ME ALLL

掩码序列: MAF SAE [mask] KE YD RRRR ME ALLL

本发明通过对大量无标签的蛋白数据进行无监督学习,学习到了丰富的蛋白质 motif 相关信息,然后应用迁移学习技术将这些模型应用到特定的有标签分类任务中,包括:对大量无标签的蛋白质序列输入蛋白质序列分词器,对蛋白质序列进行分词,以氨基酸组合残基的形式表征蛋白质序列;构建掩码语言预训练模型(MLM),在输入的分词后蛋白质序列中进行随机的氨基酸残基组合掩盖,模型在推断被掩盖的氨基酸残基组合时学习到蛋白质序列的上下文信息和表述规则,帮助模型学习到关键的残基特征。

1.5、微调蛋白质表征预训练模型

1.5.1、将生成的带有标签的可以与 MHC-I/HLA-I 结合的阳性肽段以及通过阳性肽段随机生成的阴性肽段,分别标注标签为 1/0,值得注意的是,这些阳性数据是通过实验获得的,这就意味着只能拿到有限的带标签的训练数据,这也是先使用大量的无标签蛋白质数据做预训练模型的原因。

1.5.2 将 1.5.1 中的标签的数据按照一定的比例划分为训练集/验证集/测试集。

1.5.3、使用 1.5.2 中的数据对生成的预训练模型进行微调,输入构建二分类模型。

1.5.4、模型结构可见图 2。

1.6、二分类模型应用

1.6.1、识别具有与 MHC-I/HLA-I 结合潜力的肽段。

1.6.2、提取蛋白质序列中的重要特征,包括:蛋白质功能性氨基酸残基组合 motif、重要作用/结合位点等。

经过上述方法,可构建从生物学特征出发,能够预测肽段与 MHC-I/HLA-I 具有结合潜力的肽段,并鉴定其重要特征。

1.7、识别突变位置对肽段与 MHC-I/HLA-I 结合和 TCR 识别潜力的影响

根据获取的肽段和 MHC-I/HLA-I 结合位点,结合野生肽段/突变肽段的突变所在位置,考虑中枢耐受机制在新抗原筛选中的影响,将 pMHC 结合和 TCR 识别类型分为四类,包括:有新抗原潜力的前 3 类,及需要从新抗原筛选中排除的最后 1 类:

第一类:野生肽为弱 MHC-I/HLA-I 结合,弱 TCR 识别,突变肽为 MHC-I/HLA-I 结合位点突变,转为强 MHC-I/HLA-I 和强 TCR 识别,在没有中枢耐受的情况下,突变发生在锚定位也可激活 TCR;

第二类:野生肽为弱 MHC-I/HLA-I 结合,弱 TCR 识别,突变肽为非 MHC-I/HLA-I 结合位点突变,转为强 MHC-I/HLA-I 和强 TCR 识别,在无中枢耐受的情况下,突变发生在非锚定位,能激活 TCR;

第三类:野生肽为强 MHC-I/HLA-I 结合,弱 TCR 识别,突变肽为非 MHC-I/HLA-I 结合位点突变,转变为强 MHC-I/HLA-I 和强 TCR 识别,在有中枢耐受的情况下,突变发生在非锚定位可以激活 TCR;

第四类:野生肽为强 MHC-I/HLA-I 结合,弱 TCR 识别,突变肽为 MHC-I/HLA-I 结合位点突变,转为强 MHC-I/HLA-I 和弱 TCR 识别。在有中枢耐受的情况下,突变发生在非锚定位,不能激活 TCR。进一步解释为:这种组合情况下,突变位点发生在 MHC-I/HLA-I 结合位点,暴露在 TCR 识别端的序列并未发生改变,进一步的此条肽段仍旧受中枢免疫耐受的影响,并不是合适的新生抗原肽。

通过上述步骤,本发明根据生物学特征和注意力分数,推断肽段和 MHC-I/HLA-I 识别时起重要作用的特征,包括:表征 MHC-I/HLA-I 序列,获取关键作用氨基酸残基组合及所在位置的方法,根据二分类模型,预测 MHC-I/HLA-I 序列中氨基酸残基组合类型及其所在的不同位置在肽段识别,结合和呈递中所起的作用,根据注意力得分获取关键位点的重要氨基酸残基组合;推断肽段关键结合位点,获取关键作用氨基酸及其所在位置。利用获得的肽段结合位点信息,根据突变所在肽段的位置,对野生型/突变型肽段与 MHC-I/HLA-I 结合以及 TCR 识别类型进行分类,引入中枢耐受对新抗原筛选的影响,根据分类结果对新抗原肽段进行进一步的剔除,获取更为准确的新抗原。

实施例二

本发明从 UniProt 蛋白质数据库中获取蛋白序列进行分词器构建;从 IEDB 获取阳

性肽与 MHC-I/HLA-I 结合序列，针对阳性序列进行阴性肽段生成；经过分词器分词后的蛋白序列构建蛋白质表征预训练模型；经过分词器分词后的 MHC-I/HLA-I 序列和肽段转化为数字表示的格式并拼接到一起，输入蛋白质表征预训练模型对其进行微调，最终获得可对肽段和 MHC-I/HLA-I 结合潜力进行预测的二分类模型；根据训练生成的模型及算法特点，提取蛋白质序列中的关键作用位点和 motif，表征 MHC-I/HLA-I 序列；利用肿瘤特异性突变在抗原肽上所处的位置和与 MHC 的结合潜力，进一步的将抗原肽进行分类，剔除掉由于受中枢免疫耐受影响而无法被 TCR 识别的抗原肽，最终获得更为准确和全面的新生抗原。具体包括以下步骤：

2.1、分词器构建及使用

使用进行训练，设置词汇表大小为 10,000，分词器效果如下：

分词前：MIINPTSDPEVSALEKKNTGRIAQIIGPVLDVTFPPGKMP.....

分词后：MII NP TTSD PE VS ALEK KN TGRI AQI IGP VLD VT FP PG KMP.....

2.2、预训练模型构建

使用自然语言模型进行蛋白质表征预训练模型构建，对分词后的蛋白质序列进行无监督训练，学习蛋白序列的语言特点及上下文联系，设置掩码比例为 15%，使用的蛋白质序列为 568,745 条，总长 sentencepiece 度 206M。

2.3、微调蛋白质表征预训练模型

2.3.1、根据与 MHC-I/HLA-I 结合的阳性肽生成对应的阴性肽，对 MHC-I/HLA-I 序列进行分词，阴阳数据条目分别标注为 0/1，阳性数据集分布如下：

表 1

阳性肽数量	阴性肽数量	HLA 种类
502,735	12,505,198	121

如表 1 所示，HLA 种类仅有 121 种，因此使用如此有限的数据集进行蛋白质特征识别是非常困难的，因此这是先构建蛋白质表征预训练模型的原因。

2.3.2、输入到蛋白质表征预训练模型中的原始数据格式示例：

E I E I C D G F,M RVT AP RTLL LLL WG AV ALT ET WAG SH SM RY FH TS VS
RPG RG EP RFI TVG YVD DT.....AGL AV ----- L AVV - VIG AV VAA VM - CRR KS
SGG KGG SY SQ AAC SD S AQG SD VSL TA -----,0

2.4、二分类模型应用

2.4.1、按照上述 2.3.2 中的数据格式输入二分类模型，可以预测肽段与 MHC-I/HLA-I 的结合潜力，输入结果如表 2 所示：

表 2

peptide	MHC	Label	prediction	probability
KAISKNVLF	MRVT.....QTA-----	1	1	[0.050, 0.950]
-	-	-	-	-

RPTRPDKARR	MAV.....CKV-----	0	0	[0.990,0.001]
------------	------------------	---	---	---------------

其中，Label 为实际标签；prediction 为模型预测标签；
probability 为模型预测结果[A,B]，A 为结果为阴性的概率，B 为结果是阳性的概率
2.4.2、评估二分类模型性能。

评估数据集为独立于模型训练所用数据集的验证集，验证集构成如表 3 所示：

表 3

阳性肽数量	阴性肽数量	HLA 种类
3,725	108,113	40

使用下述公式计算模型性能：

$$Precision = \frac{TP}{TP + FP}$$
$$Recall/TPR = \frac{TP}{TP + FN}$$
$$FPR = \frac{FP}{TN + FP}$$

TP：真阳性； TN：真阴性； FP：假阳性； FN：假阴性；

AUC 是真阳性率 (TPR) 与假阳性率 (FPR) 的曲线。

与常见的 NetMHCpan4.1 软件做比较，进行模型性能评估，结果如表 4 所示：

表 4

Type	NetMHCpan4.1	LKMHCpan
AUC	0.77	0.85
Precision	0.75	0.84
Recall	0.79	0.85

2.4.3、提取 MHC-I/HLA-I 结合位点，获取关键作用氨基酸残基位置的方法，从生物学角度解释模型预测性能，根据不同位点的注意力得分绘制热图，颜色越亮，更有可能是肽段与 MHC 的关键结合位点，也侧面证明了模型的良好性能，可以用于后面的 MHC-I/HLA-I 序列的关键位点特征提取。

2.4.4、表征蛋白质序列，获取关键作用氨基酸残基组合及所在位置的方法，从生物学角度解释模型预测性能。绘制 HLA-A/B/C 序列上每个位置的注意力得分热图，推断出 MHC-I/HLA-I 序列总的关键结合位点，帮助更好的表征 MHC-I/HLA-I 序列，与常见的 LKMHCpan 或者 MHCflurry 提供的热点位 (31, 33, 48, 69, 83, 86, 87, 90, 91, 93, 94, 97, 98, 100, 101, 104, 105, 108, 115, 119, 121, 123, 126, 138, 140, 142, 167, 171, 174,176, 180, 182, 183, 187, 191, 195, 223)更为全面和更具代表性，关注热点的功能性 motif 序列，提取涵盖上述位点之外更多的关键位点，例如 39,235 等位点，更符合蛋白序列所蕴含的功能意义。

2.5、增加肽段与 MHC-I/HLA-I 结合及 TCR 识别的分类信息，更为详细和全面的推动新抗原序列识别。

根据上述分析后获得肽段与 MHC-I/HLA-I 结合锚定位置，可根据野生肽/突变肽的

结合潜力及肿瘤特异性突变位置，将肽段类型分为 4 组，如下表 5 所示：

表 5

类别	突变发生位置	野生型肽段	突变型肽段	是否适合被选为新抗原
1	MHC 结合位点	弱 MHC 结合&弱 TCR 识别	MHC 强结合&TCR 强识别	是
2	非 MHC 结合位点	弱 MHC 结合&弱 TCR 识别	MHC 强结合&TCR 强识别	是
3	非 MHC 结合位点	强 MHC 结合&弱 TCR 识别	MHC 强结合&TCR 强识别	是
4	MHC 结合位点	强 MHC 结合&弱 TCR 识别	MHC 强结合&TCR 弱识别	否

在新生抗原筛选中剔除掉类别 4 产生的抗原肽，用于提升新生抗原预测的准确性。

实施例三

步骤 3.1 至 3.4 同实施例一中 1.1 至 1.4。

3.5、微调蛋白质表征预训练模型

3.5.1、将生成的带有标签的可以与 MHC-I/HLA-I 结合的阳性肽段以及通过阳性肽段随机生成的阴性肽段，分别标注标签为 1/0，值得注意的是，这些阳性数据是通过实验获得的，这就意味着只能拿到有限的带标签的训练数据，这也是先使用大量的无标签蛋白质数据做预训练模型的原因。

3.5.2 将 3.5.1 中的带有标签的蛋白质分词器分词的 MHC-I/HLA-I 序列和肽段序列进行数字化表示并拼接，拼接策略如下：

通过肽段和其所在的蛋白质序列计算蛋白质序列裂解可能性；

通过肽段和 MHC-I/HLA-I 计算其结合稳定性，及结合后呈递能力；

将数字表示的白质序列裂解可能性、肽段和 MHC-I/HLA-I 结合稳定性以及肽段和 MHC-I/HLA-I 结合后呈递能力，与数字化表示的蛋白质分词器分词的 MHC-I/HLA-I 序列和肽段序列进行拼接。

3.5.3、将 3.5.2 中的标签的数据按照一定的比例划分为训练集/验证集/测试集。

3.5.4、使用 3.5.3 中的数据对生成的预训练模型进行微调，输入构建二分类模型。

3.5.5、模型结构可见图 3。

步骤 3.6 至 3.7 同实施例一中 1.6 至 1.7。

本发明根据生物学特征和注意力分数，推断肽段和 MHC-I/HLA-I 识别时起重要作用的特征，包括：表征 MHC-I/HLA-I 序列，获取关键作用氨基酸残基组合及所在位置的方法，根据二分类模型，预测 MHC-I/HLA-I 序列中氨基酸残基组合类型及其所在的不同位置在肽段识别，结合和呈递中所起的作用，根据注意力得分获取关键位点的重要氨基酸残基组合；推断肽段关键结合位点，获取关键作用氨基酸及其所在位置。利用获得的肽段结合位点信息，根据突变所在肽段的位置，对野生型/突变型肽段与 MHC-I/HLA-I 结合以及 TCR 识别类型进行分类，引入 pMHC 结合稳定性、pMHC 呈递能力和中枢耐受对新抗原筛选的影响，根据分类结果对新抗原肽段进行进一步的剔除，获取更为准确的新抗原。

实施例四

本发明从 UniProt 蛋白质数据库中获取蛋白序列进行分词器构建；从 IEDB 获取阳性肽与 MHC-I/HLA-I 结合序列，针对阳性序列进行阴性肽段生成；经过分词器分词后的蛋白序列构建蛋白质表征预训练模型；经过分词器分词后的 MHC-I/HLA-I 序列和肽段转化为数字表示的格式；通过肽段和其所在的蛋白质序列计算蛋白质序列裂解可能性；通过肽段和 MHC-I/HLA-I 计算其结合稳定性，及结合后呈递可能性；将上述数字表示 MHC-I/HLA-I 序列和肽段及其对应的肽段裂解可能性、pMHC 结合稳定性及 pMHC 呈递可能性拼接到一起，拼接方式为 torch.cat()，输入蛋白质表征预训练模型对其进行微调，继承原本的预训练模型并增加全连接层和分类层，最终获得可对肽段和 MHC-I/HLA-I 结合潜力进行预测的二分类模型；根据训练生成的模型及算法特点，提取蛋白质序列中的关键作用位点和 motif，表征 MHC-I/HLA-I 序列；利用肿瘤特异性突变在抗原肽上所处的位置和与 MHC 的结合潜力，进一步的将抗原肽进行分类，剔除掉由于受中枢免疫耐受影响而无法被 TCR 识别的抗原肽，最终获得更为准确和全面的新生抗原。具体包括以下步骤：

步骤 4.1 至 4.2 同实施例二中 2.1 至 2.2

4.3、微调蛋白质表征预训练模型

4.3.1、根据与 MHC-I/HLA-I 结合的阳性肽生成对应的阴性肽，对 MHC-I/HLA-I 序列进行分词，阴阳数据条目分别标注为 0/1，阳性数据集分布如下表 6 所示：

表 6

阳性肽数量	阴性肽数量	HLA 种类
502,735	12,505,198	121

4.3.2、计算影响肽段与 MHC-I/HLA-I 结合的特征，包括：肽段的裂解可能性、pMHC 稳定性、pMHC 结合后呈递可能性，例如可以通过如下方式计算：

使用软件 pepsickle-v0-2-2,输入肽段序列,输出每个位点裂解可能性得分,肽段的裂解可能性为肽段的 C-/N-端氨基酸的裂解得分的平均值，范围[0,1]，得分越大裂解可能性越高；

使用软件 LKMHCstablepan-v1，输入肽段序列及 MHC-I/HLA-I 序列，计算 pMHC 结合稳定性，输出 pMHC 结合稳定性得分，得分越高说明越稳定；

使用软件 DeepTAP-v1.0.1，输入肽段序列，计算 pMHC 结合后呈递至细胞表面可能性，输出 pMHC 结合后呈递至细胞表面可能性评分等级，范围为[0,1]，评分越大呈递可能性越高。

4.3.3、输入到蛋白质表征预训练模型中的原始数据格式示例：

E I E I C D G F,M RVT AP RTLL LLL WG AV ALT ET WAG SH SM RY FH TS VS
RPG RG EP RFI TVG YVD DT.....AGL AV ----- L AVV - VIG AV VAA VM - CRR KS
SGG KGG SY SQ AAC SD S AQG SD VSL TA -----,0.057,0.20,0.12,0 （依次对应：

肽段序列、MHC-I/HLA-I 序列、肽段裂解可能性，pMHC 结合稳定性得分，pMHC 结合后呈递可能性评分等级，阴阳性)

将经过分词器分词后的 MHC-I/HLA-I 序列和肽段转化为数字表示的格式的序列后，与肽段的裂解可能性得分、pMHC 的结合稳定性得分、pMHC 呈递可能性得分沿着指定的维度拼接，输入蛋白质表征预训练模型进行模型训练。

4.4、二分类模型应用

4.4.1、按照 4.3.3 中的数据格式输入二分类模型，可以预测肽段与 MHC-I/HLA-I 的结合潜力，输入结果如表 7 所示：

表 7

peptide	MHC	Label	prediction	probability
KAISKNVLF	MRVT.....QTA-----	1	1	[0.0320, 0.9681]
RPTRPKARR	MAV.....CKV-----	0	0	[0.9995, 0.0005]

其中，Label 为实际标签；prediction 为模型预测标签；probability 为模型预测结果 [A,B]，A 为结果为阴性的概率，B 为结果是阳性的概率

4.4.2、评估二分类模型性

评估数据集同实施例二中 2.4.2 一致：

计算模型性能的公式同实施例二中 2.4.2 一致：

与常见的 NetMHCpan4.1 软件，及实施例二构建的二分类模型相比，进行模型性能评估，如表 8 所示：

表 8

Type	NetMHCpan4.1	LKMHCpan (实施例二)	LKMHCpan (实施例四)
AUC	0.77	0.85	0.89
Precision	0.75	0.84	0.91
Recall	0.79	0.85	0.86

步骤 4.4.3 至 4.4.4 同实施例二中 2.4.3 至 2.4.4。

步骤 4.5 同实施例二中 2.5。

以上参照附图说明了本发明的优选实施例，并非因此局限本发明的权利范围。本领域技术人员不脱离本发明的范围和实质内所作的任何修改、等同替换和改进，均应在本发明的权利范围之内。

权利要求书

1.一种用于鉴定与 MHC-I/HLA-I 类结合及 TCR 识别肽段的方法，其特征在于，包括：

获取蛋白质序列、MHC-I/HLA-I 序列和肽段序列；

利用所述蛋白质序列训练蛋白质序列分词器；

利用经过所述蛋白质序列分词器进行分词处理后的蛋白质序列进行蛋白质表征模型训练，得到训练后的蛋白质表征预训练模型；

利用经过所述蛋白质序列分词器进行分词处理后的 MHC-I/HLA-I 序列和肽段序列对所述训练后的蛋白质表征预训练模型进行微调，得到二分类模型；以及

利用所述二分类模型对目标野生型/突变型肽段进行鉴定处理，对突变型肽段进行分类，获取与 TCR 识别正相关的候选新生抗原突变肽；

其中，所述 MHC-I 是指主要组织相容性复合物 I 类；所述 HLA-I 是指人类中的主要组织相容性复合物 I 类；所述 TCR 是指 T 细胞受体。

2.根据权利要求 1 所述的方法，其特征在于，所述肽段序列包括阳性肽段序列和阴性肽段序列，其中，所述阳性肽段序列为与 MHC-I/HLA-I 结合的肽段，标签为 1；所述阴性肽段序列为根据阳性肽段序列特征在蛋白序列中随机生成的肽段序列，标签为 0。

3.根据权利要求 2 所述的方法，其特征在于，所述蛋白质序列包括从蛋白质数据库中收集蛋白质序列；相应地，利用所收集的蛋白质序列进行高频重复氨基酸残基组合 motif 的分词器训练，获得蛋白质序列分词器。

4.根据权利要求 3 所述的方法，其特征在于，所述利用经过所述蛋白质序列分词器进行分词处理后的蛋白质序列进行蛋白质表征模型训练，得到训练后的蛋白质表征预训练模型，包括：

将无标签的蛋白质序列输入至蛋白质序列分词器中进行分词处理，得到氨基酸功能残基组合 motif 分割后的蛋白质序列；

将所述氨基酸功能残基组合 motif 分割后的蛋白质序列转换为数字表示的蛋白质序列；以及

利用所述数字表示的蛋白质序列进行蛋白质表征模型训练，得到训练后的蛋白质表征预训练模型。

5.根据权利要求 4 所述的方法，其特征在于，所述利用经过所述蛋白质序列分词器进行分词处理后的 MHC-I/HLA-I 序列和肽段序列对所述训练后的蛋白质表征预训练模型进行微调，得到二分类模型包括：

将带标签的 MHC-I/HLA-I 序列输入至蛋白质序列分词器中进行分词处理，得到氨基酸功能残基组合 motif 分割后的 MHC-I/HLA-I 序列，并将所述氨基酸功能残基组合 motif 分割后的 MHC-I/HLA-I 序列转换为数字表示的 MHC-I/HLA-I 序列；

将带标签的肽段序列转换为数字表示的肽段序列；以及

通过将所述转换为数字表示的 MHC-I/HLA-I 序列和所述转换为数字表示的肽段序列进行拼接处理，得到拼接后的序列，并利用所述拼接后的序列对所述训练后的蛋白质表征预训练模型进行微调，得到二分类模型。

6.根据权利要求 5 所述的方法，其特征在于，所述拼接处理进一步包括将所述转换为数字表示的 MHC-I/HLA-I 序列、转换为数字表示的肽段序列、转换为数字表示的与肽段和 MHC-I/HLA-I 结合相关的其他特征，所述其他特征选自由所述蛋白质序列的裂解可能性、MHC-I/HLA-I 与肽段的结合稳定性，和 MHC-I/HLA-I 与肽段复合物呈递可能性中的一种或多种。

7.根据权利要求 6 所述的方法，其特征在于，在所述微调的步骤之前，所述方法进一步包括利用所述蛋白质序列、MHC-I/HLA-I 序列和肽段序列计算所述与肽段和 MHC-I/HLA-I 结合相关的其他特征。

8.根据权利要求 5 所述的方法，其特征在于，所述利用所述二分类模型对目标野生型/突变型肽段进行鉴定处理包括：

将目标野生型/突变型肽段输入至所述二分类模型中，输出预测为阳性的突变型肽段每个位点的注意力得分，以及 MHC-I/HLA-I 序列的氨基酸功能残基组合 motif 的注意力得分；以及

根据所述阳性突变型肽段每个位点的注意力得分以及 MHC-I/HLA-I 序列 motif 的注意力得分，判断在 MHC-I/HLA-I 序列与阳性肽段结合时的锚定位以及 MHC-I/HLA-I 序列的关键表征位点，并利用所述 MHC-I/HLA-I 序列与肽段结合时的锚定位点和肽段的突变位点对阳性突变型肽段进行分类处理。

9.根据权利要求 8 所述的方法，其特征在于，所述利用所述 MHC-I/HLA-I 序列与阳性突变型肽段结合时的锚定位以及相对于野生型肽段的突变位进行分类处理包括：

根据所述 MHC-I/HLA-I 序列与肽段结合时的锚定位点，并结合突变所在位置进行阳性突变型肽段分类处理，并根据分类结果判断是否适合被选为 TCR 识别正相关的候选新生抗原突变肽。

10.根据权利要求 9 所述的方法，其特征在于，所述根据分类结果判断是否适合被选为 TCR 识别正相关的候选新生抗原突变肽包括：

当野生型肽段为弱 MHC-I/HLA-I 结合且弱 TCR 识别，突变型肽段为 MHC-I/HLA-I 锚定位点突变，转为强 MHC-I/HLA-I 和强 TCR 识别时，则判断能被选为新生抗原突变肽；

当野生型肽段为弱 MHC-I/HLA-I 结合且弱 TCR 识别，突变型肽段为非 MHC-I/HLA-I 锚定位点突变，转为强 MHC-I/HLA-I 和强 TCR 识别时，则判断能被选为新生抗原突变肽；

当野生型肽段为强 MHC-I/HLA-I 结合且弱 TCR 识别, 突变型肽段为非 MHC-I/HLA-I 锚定位点突变, 转变为强 MHC-I/HLA-I 和强 TCR 识别时, 则判断能被选为新生抗原突变肽;

当野生型肽段为强 MHC-I/HLA-I 结合且弱 TCR 识别, 突变型肽段为 MHC-I/HLA-I 锚定位点突变, 转为强 MHC-I/HLA-I 和弱 TCR 识别时, 则判断不能被选为新生抗原突变肽。

11. 一种电子设备, 其特征在于, 包括: 存储器; 处理器; 以及计算机程序; 其中, 所述计算机程序存储在所述存储器中, 并被配置为由所述处理器执行以实现如权利要求 1-10 任一项所述的方法。

12. 一种计算机可读存储介质, 其特征在于, 其上存储有计算机程序; 所述计算机程序被处理器执行以实现如权利要求 1-10 任一项所述的方法。