(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113762417 B
(45)授权公告日 2022.05.27

(21)申请号 202111204491.8
(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 113762417 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 方樹楷 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.
G06K 9/62(2022.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)
审查员 刘昶忻

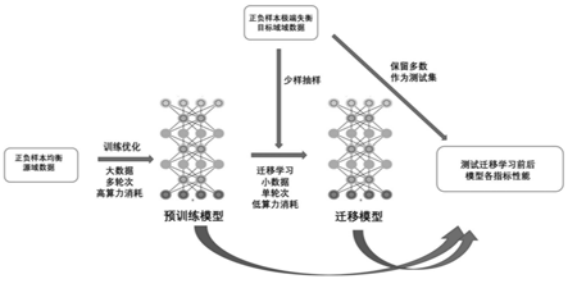
权利要求书2页 说明书9页 附图1页

(54)发明名称

基于深度迁移的对HLA抗原呈递预测系统的增强方法

(57)摘要

本发明提出了一种基于深度迁移的对HLA抗原呈递预测系统的增强方法，包括：1)使用全局最大差异打分矩阵生成不同比例的负样本训练集：正负样本均衡的源域数据集，正负样本失衡的目标数据集2)采用多种不同的深度神经网络来编码已知序列信息、多模态特征融合等手段，在正负样本比均衡的源域数据上得到预训练模型3)通过深度迁移方法，将预训练模型迁移至正负样本比极端失衡的目标数据集4)提出创新的“严格准确率(strict PPV)”指标。同于以往其他基于单一数据集与单一人工智能模型的MHC预测方法，本发明能高效地融合多模态信息，快速部署迁移到不用的数据集上，节省了在新环境与数据上重新训练模型的算力与时间成本。



CN 113762417 B

1.一种基于深度迁移的对HLA抗原呈递预测系统的增强方法,其特征在于,包括以下步骤:

S1、特征选择与归一化处理,构建原始领域数据作为源域数据集;

S1中特征选择的不同,选定相应的归一化处理方案,以获取格式、维度统一,便于融合的特征向量,具体为:

-长序列特征,使用随机矩阵将其每个氨基酸编码到可学习的到隐空间,再利用长短记忆循环神经网络进行处理;

-短序列特征,利用独热方法进行编码,编码后的序列送入多层感知机网络模型进行变换;

-向量特征,采用主成分分解PCA进行编码,将所有数据的向量形式的特征组合成特征矩阵,应用主成分分解进行矩阵分解;根据隐嵌入维度选择特定数目的矩阵特征向量作编码变换;

-标量特征,采用多维尺度放缩,高斯核方法进行编码:将所有数据的标量形式的特征作为高斯核的输入,得到高斯核的协方差矩阵;将矩阵的各列进行多维尺度放缩,得到编码变换的特征向量;

S2、特征融合与训练求解预训练模型;

S3、构建专用的极端不平衡的迁移目标领域数据作为目标域数据集;

S4、将S2中得到的预训练模型,利用深度迁移方法迁移到S3中的目标领域数据,以构建深度迁移自适应优化模型;

S4中构建的深度迁移自适应优化模型:

$$\min_{W'} \sum_{i=1}^{N_1} \mathrm{loss}_S\left(f'_{W'}, D_S^i, Y_S^i\right) + \lambda \sum_{i=1}^{N_2} \mathrm{loss}_C\left(f'_{W'}, D_C^i, Y_C^i\right)$$

式中,$f'_{W'}$是含学习参数的待迁移的预测模型;

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;

$\mathrm{loss}_S, \mathrm{loss}_C$分别表示在预训练阶段与模型迁移自适应阶段的目标损失函数;λ表示赋予模型迁移自适应阶段的目标损失函数的权重;

$(D_S^i, Y_S^i), (D_C^i, Y_C^i)$分别表示S1中构建的原始领域数据集与S3中构建的目标领域数据集上的训练数据特征与是否呈递结合的真实值;

N_1, N_2分别表示S1中构建的原始领域数据集与S3中构建的目标领域数据集上的训练样本数量;

在对模型进行优化后,将相关参数以结构化方法保存为自适应后的深度迁移自适应优化模型;

S5、使用深度迁移自适应优化模型,在目标领域数据集上进行HLA抗原呈递预测。

2.根据权利要求1所述的方法,其特征在于S1中构建的原始领域数据集是正负样本数量比均衡。

3.根据权利要求1所述的方法，其特征在于S1中构建的原始领域数据集时，使用窗口滑动的方法，根据预设的参数阈值，生成阴性序列并使用全局差异打分矩阵筛选生成的序列片段，获得非随机的阴性候选训练集。

4.根据权利要求1所述的方法，其特征在于S2中待融合特征选择为：多肽序列特征、上下游序列特征、呈递亲和力特征。

5.根据权利要求4所述的方法，其特征在于S2中：

多肽序列特征通过以下方法获得标准特征：对于给定多肽肽链氨基酸序列，使用随机矩阵将其每个氨基酸编码到可学习的到隐空间，再利用长短记忆循环神经网络进行处理得到多肽序列特征；进行随机矩阵编码映射后，根据所有数据中最长肽链序列的长度进行补长，以保证编码与映射模型的参数保持一致；

上下游序列特征通过以下方法获得标准特征：对于给定基因上下游肽链，利用独热方法进行编码，编码后的上下游序列进行裁剪得到定长的序列，此编码序列送入多层感知机网络模型进行变换，提取特征作为上下游序列特征；

呈递亲和力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。

6.根据权利要求1所述的方法，其特征在于S2中构建的预训练优化模型：

$$\min_W \sum_{n=1}^{N} -w_n \left[y_n \cdot \log \sigma \left(f_W(x_n) \right) + (1 - y_n) \cdot \log \left(1 - \sigma \left(f_W(x_n) \right) \right) \right]$$

式中，f_W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；w_n表示对不同样本的损失函数所赋予权重，N表示样本总数；

x_n表示输入的特定数据，y_n是训练数据中是否呈递结合的真实值，σ是S逻辑函数，非简单加和，模型公式捕捉了潜在的复杂关系；

在对模型进行优化后，将相关参数以结构化方法保存为预训练模型。

7.根据权利要求1所述的方法，其特征在于，S3 中生产阴性候选数据集后，根据不同策略构建目标领域数据；目标领域数据集为阴性样本数量远多于阳性样本数量，以模拟真实预测环境中阴性样本远多于阳性样本的情况。

8.根据权利要求1所述的方法，其特征在于所述的深度迁移自适应优化模型，根据预训练模型大小与数据规模，选择优化预训练模型中所有可训练参数的全局优化，或仅进行神经网络模型中的最后两层的选顶层优化。

9.根据权利要求1所述的方法，其特征在于S2，S4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，获得预训练预测模型f_W与迁移预测模型$f'_{W'}$。

10.根据权利要求1所述的方法，其特征在于S3所构建的极端不平衡数据中划分出单独的一批数据，用于在S5中验证深度迁移自适应优化模型对于目标领域数据对的预测效果。

基于深度迁移的对HLA抗原呈递预测系统的增强方法

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迁移的对HLA新生抗原呈递预测系统的增强方法。

背景技术

[0002] 人类白细胞抗原(human leukocyte antigen，缩写为HLA)，是编码人类的主要组织相容性复合体(MHC)的基因，与人类的免疫系统功能密切相关。MHC主要分两类，第一类MHC处理细胞内部被分解后的蛋白质(例如病毒的)、第二类当外部入侵者经过胞吞并利用溶酶体处理后形成碎片，MHC再跟这些碎片结合，并呈现在细胞表面上供T细胞所辨识。它们与人类的免疫系统功能密切相关。其中部分基因编码细胞表面抗原，成为每个人的细胞不可混淆的“特征”，是免疫系统区分本身和异体物质的基础。利用HLA呈递原理的癌症疫苗是当今医学与药物学的热点问题。

[0003] 预测个性化HLA新生抗原(Neoantigen)疫苗的主要步骤如下：

[0004] (1)鉴定和确认在患者肿瘤中表达的具有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、读码框偏移、。

[0005] (2)使用主要组织相容性复合物(MHC)I和II类表位预测算法筛选，分析和鉴定具有最高抗原呈递可能性的突变。

[0006] (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。

[0007] 步骤(2)中涉及的HLA新生抗原呈递预测问题，目是Neoantigen疫苗开发的核心点。随着人工智能技术在生物信息学中的广泛应用，已经有领域外内学者开始尝试利用数据驱动的机器学习方法，快速发现、预测与筛选可用的新生免疫抗原靶点。其中代表性的工作与技术有，美国杜克大学的NetHMCPan系列的工作(参考文献:Jurtz,Vanessa,et al."NetMHCpan-4.0:improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data."The Journal of Immunology 199.9(2017):3360-3368,Reynisson,Birkir,et al."NetMHCpan-4.1and NetMHCIIpan-4.0:improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data."Nucleic acids research 48.W1(2020):W449-W454.)采用多层感知机模型预测洗脱配体与MHC抗原结合的亲和力指数，为后续一系列预测模型提供了新的数据特征;丹麦科技大学团队的工作(参考文献:Reynisson,Birkir,et al."Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data."Journal of proteome research 19.6(2020):2304-2315.)基于洗脱配体数据与基序列的反卷积模型集合来预测MHC II类抗原的预测;美国北卡教堂山大学的团队的工作(参考文献:Smith C C,Chai S,Washington A R,et al.Machine-learning prediction of tumor antigen immunogenicity in the

selection of therapeutic epitopes[J].Cancer immunology research,2019,7(10):1591-1604.)基于免疫原多表位选择的MHC抗原预测模型。

[0008] 以上的介绍的工作,其贡献主要在于将机器学习模型应用于提前构建好的特定MHC数据集。但在实际的抗肿瘤疫苗开发应用场景中,决大多数现成的基于质谱(MS)鉴定数据集与真正的人类白细胞抗原(HLA)匹配的表位数量仍然相对较少,且数据来源多样,数据分布差异大,这使得建立一套可靠、稳健、可复用的HLA抗原预测系统十分困难。此类困难不仅是出现在HLA抗原预测中,在更广义的生物信息学、临床医学中,类似的数据获取成本高与稳健模型训练难,也是引入机器学习模型时十分突出的问题。

[0009] 因此有学者提出利用迁徙学习等前沿人工智能方法,在以较低的模型训练与数据收集成本,快速将已有的模型部署到新的场景中。迁移学习是运用已存有的知识对不同但相关领域问题进行求解的一种新的机器学习方法。一些将该方法运用在MHC预测处理领域的代表性工作有:南方科技大学的学者尝试用迁移学习的范式训练模型来学习混合等位基因特异性表位的共同特征(参考文献:Hu,Wei Peng,You Ping Li,and Xiu Qing Zhang."MHC-I epitope presentation prediction based on transfer learning."Yi Chuan=Hereditas 41.11(2019):1041-1049.)英国学者基于使用来自癌症患者的免疫肽组源域数据,利用迁移学习将相关天然呈递肽的理化特性进行编码,以此来预测肿瘤抗原肽的HLA呈递(相关文献:Ng FS,Vandenberghe M,Portella G,Cayatte C,Qu X,Hanabuchi S,Landry A,Chaerkady R,Yu W,Collepardo-Guevara R,Sidders B.MINERVA:Learning the Rules of HLA Class I Peptide Presentation in Tumors with Convolutional Neural Networks and Transfer Learning.Available at SSRN 3704016.)。但总体而言,相关的研究还是比较少。

[0010] 该领域主流方法模型落地的另一个挑战就是真实情况下极端不平衡的数据比。在训练模型过程中通常需要选择正负样本接近的数据集,以确保模型训练过程平稳。但真实场景下负样本(未呈递)远多于正样本。这进一步提高了评价模型真实性能的难度。目前学术界对此有少许讨论(参考文献:Schneider M,Wang L,Marr C.Evaluation of domain adaptation approaches for robust classification of heterogeneous biological data sets.InInternationalConference on Artificial Neural Networks 2019Sep 17(pp.673-686).Springer,Cham.),但该问题并未得到学术界的广泛重视与讨论。

发明内容

[0011] 针对背景技术中提到的主流HLA预测存在的两个主要问题:1.模型在异质数据上迁徙时效果差,2.现有指标难以衡量模型在极端正负比下的真实性能,本发明提出了一种一种基于深度迁移的对HLA抗原呈递预测系统的增强方法。

[0012] 本发明首先公开了一种基于深度迁移的对HLA抗原呈递预测系统的增强方法,包括以下步骤:

[0013] S1、特征选择与归一化处理,构建原始领域数据作为源域数据集;

[0014] S2、特征融合与训练求解预训练模型;

[0015] S3、构建专用的极端不平衡的迁移目标领域数据作为目标域数据集;

[0016] S4、将S2中得到的预训练模型,利用深度迁移方法迁移到S3中的目标领域数据,以

构建深度迁移自适应优化模型；

[0017] S5、使用深度迁移自适应优化模型，在目标领域数据集上进行HLA抗原呈递预测。

[0018] 优选的，根据S1中特征选择的不同，选定相应的归一化处理方案，以获取格式、维度统一，便于融合的特征向量，具体为：

[0019] -长序列特征，使用随机矩阵将其每个氨基酸编码到可学习的到隐空间，再利用长短记忆循环神经网络进行处理；

[0020] -短序列特征，利用独热方法进行编码，编码后的序列送入多层感知机网络模型进行变换；

[0021] -向量特征，采用主成分分解PCA进行编码，将所有数据的向量形式的特征组合成特征矩阵，应用主成分分解进行矩阵分解；根据隐嵌入维度选择特定数目的矩阵特征向量作编码变换；

[0022] -标量特征，采用多维尺度放缩，高斯核方法进行编码：将所有数据的标量形式的特征作为高斯核的输入，得到高斯核的协方差矩阵；将矩阵的各列进行多维尺度放缩，得到编码变换的特征向量。

[0023] 优选的，S1中构建的原始领域数据集是正负样本数量比均衡(阳性样本数量：阴性样本数量＝1：1)，以最好地适应预训练模型的结构与优化方法。

[0024] 优选的，S1中构建的原始领域数据集时，使用窗口滑动的方法，根据预设的参数阈值，生成阴性序列并使用全局差异打分矩阵筛选生成的序列片段，获得非随机的阴性候选训练集。

[0025] 优选的，S2中待融合特征选择为：多肽序列特征、上下游序列特征、呈递亲和力特征。

[0026] 优选的，S2中：

[0027] 多肽序列特征通过以下方法获得标准特征：对于给定多肽肽链氨基酸序列，使用随机矩阵将其每个氨基酸编码到可学习的到隐空间，再利用长短记忆循环神经网络进行处理得到多肽序列特征；进行随机矩阵编码映射后，根据所有数据中最长肽链序列的长度进行补长，以保证编码与映射模型的参数保持一致；

[0028] 上下游序列特征通过以下方法获得标准特征：对于给定基因上下游肽链，利用独热方法进行编码，编码后的上下游序列进行裁剪得到定长的序列，此编码序列送入多层感知机网络模型进行变换，提取特征作为上下游序列特征；

[0029] 呈递亲和力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。

[0030] 优选的，S2中构建的预训练优化模型：

[0031] $$\min_W \sum_{n-1}^{N} -w_n \left[y_n \cdot \log \sigma \left(f_W(x_n) \right) + (1 - y_n) \cdot \log \left(1 - \sigma \left(f_W(x_n) \right) \right) \right]$$

[0032] 式中，f_W是含可学习参数的预测模型；

[0033] 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；w_n表示对不同样本的损失函数所赋予权重，N表示样本总数；

[0034] x_n表示输入的特定数据，y_n是训练数据中是否呈递结合的真实值，o是S逻辑函数，

非简单加和,模型公式捕捉了潜在的复杂关系;

[0035] 在对模型进行优化后,将相关参数以结构化方法保存为预训练模型。

[0036] 优选的,S3中生产阴性候选数据集后,根据不同策略构建目标领域数据;目标领域数据集为阴性样本数量远多于阳性样本数量,以模拟真实预测环境中阴性样本远多于阳性样本的情况。

[0037] 优选的,S4中构建的深度迁移自适应优化模型:

[0038] $$\min_{W'} \sum_{i=1}^{N_1} \mathrm{loss}_S\left(f'_{W'}, D_S^i, Y_S^i\right) + \lambda \sum_{i=1}^{N_2} \mathrm{loss}_C\left(f'_{W'}, D_C^i, Y_C^i\right)$$

[0039] 式中,fW′是含学习参数的待迁移的预测模型;

[0040] 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;

[0041] loss_S,loss_C分别表示在预训练阶段与模型迁移自适应阶段的目标损失函数;λ表示赋予模型迁移自适应阶段的目标损失函数的权重;

[0042] (D_S^i, Y_S^i),(D_C^i, Y_C^i)分别表示S1中构建的原始领域数据集与S3中构建的目标领域数据集上的训练数据特征与是否呈递结合的真实值;

[0043] N_1,N_2分别表示S1中构建的原始领域数据集与S3中构建的目标领域数据集上的训练样本数量;

[0044] 在对模型进行优化后,将相关参数以结构化方法保存为自适应后的深度迁移自适应优化模型。

[0045] 优选的,所述的深度迁移自适应优化模型,根据预训练模型大小与数据规模,选择优化预训练模型中所有可训练参数的全局优化,或仅进行神经网络模型中的最后两层的选顶层优化。

[0046] 优选的,S2,S4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,获得预训练预测模型f_W与迁移预测模型$f_{W'}$

[0047] 优选的,S3所构建的极端不平衡数据中划分出单独的一批数据,用于在S5中验证深度迁移自适应优化模型对于目标领域数据对的预测效果。

[0048] 本发明的有益效果

[0049] 本发明提出了一种全新的基于深度迁移的对HLA新生抗原呈递预测系统的增强方法,包括:1)使用全局最大差异打分矩阵生成不同比例的负样本训练集:正负样本均衡的源域数据集,正负样本失衡的目标数据集2)采用多种不同的深度神经网络来编码已知序列信息、多模态特征融合等手段,在正负样本比均衡的源域数据上得到预训练模型3)通过深度迁移方法,将预训练模型迁移至正负样本比极端失衡的目标数据集4)提出创新的“严格准确率(strict PPV)”指标。同于以往其他基于单一数据集与单一人工智能模型的MHC预测方法,本发明能高效地融合多模态信息,快速部署迁移到不用的数据集上,节省了在新环境与数据上重新训练模型的算力与时间成本。

[0050] 基于本申请提出的多模态特征融合预测,并非传统单一的加和,捕捉了多特征之间潜在的复杂关系。

[0051] 基于正负样本比均衡的源域数据的训练,保证了预训练模型收敛的平稳性,以及

模型学习到了多模态特征的隐嵌入表达。

[0052] 基于正负样本比极端失衡的目标数据的深度迁移,保证了迁移后模型在真实环境下的可靠性与可复用性。

附图说明

[0053] 图1基于深度迁移的对HLA新生抗原呈递预测模型进行迁移与增强方法图

[0054] 图2基于深度迁移的对HLA新生抗原呈递预测系统的增强方法和评价系统总流程图

具体实施方式

[0055] 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:

[0056] 如图2所示,本发明提出的于深度迁移的对HLA新生抗原呈递预测系统的增强和评价方法系统分为为三部分,下面针对数据集构建,模型优化迁移,与模型测试三部分进行详细阐述。

[0057] (a)数据集构建

[0058] 首先,该模块为根据公开数据库资源,特定文献,收集特定HLA新生抗原呈递的多肽肽链、上下游等数据对,亲和力指数等数据元组,以及配套的数据处理、标准化流程,具体包括:

[0059] I.给定特定蛋白质,成功表达呈递的特定HLA新生抗原的特定多态氨基酸序列。

[0060] II.该蛋白质对应的上下游各6个、共12个氨基酸长度的上下游相关序列。

[0061] III.根据I中蛋白质、多肽数据对,从一系列专业计算工具(参考文献:Jurtz,Vanessa,et al."NetMHCpan-4.0:improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data."The Journal of Immunology 199.9(2017):3360-3368.)得到的呈递表达数据对的亲和力指数(affinity score)以及相关特征。

[0062] 具体而言,我们参考的公开数据源、文献资源有主要有MARIA(参考文献:Chen,Binbin,et al."Predicting HLA class II antigen presentation through integrated deep learning."Nature biotechnology 37.11(2019):1332-1343.NetMHCpan系列数据,(参考文献:Reynisson,Birkir,et al."NetMHCpan-4.1and NetMHCIIpan-4.0:improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data."Nucleic acids research 48.W1(2020):W449-W454.),图宾根大学公开MHC配对数据(参考文献:Rammensee,H-G.,et al."SYFPEITHI:database for MHC ligands and peptide motifs."Immunogenetics 50.3(1999):213-219.)。

[0063] 基于以上方法收集来自多个数据源的HLA新生抗原成功表达呈递的正样本后,开始构建训练机器学习所用的数据集,具体的,先利用全局最大差异打分矩阵生成负样本并构建源域数据集与目标域数据集。其原理为:一般认为被呈递的肽段和正常肽段序列相似度和抗原呈递及免疫原性有一定的负相关关系,因此我们使用了全局最大差异打分矩阵的方式生成序列相似度最低肽段作为训练集的阴性样本集。生成的具体方法是,先确定正负

样本比例,再使用窗口在的序列上按顺序滑动,将所有产生的序列使用BioPython序列比对软件包进行多序列比对,并且使用冒泡法保留特定个序列相似度最低的阴性序列作为阴性训练集。

[0064] 在真实生产环境中,HLA新生抗原表达呈递失败的概率远高于其成功的概率,所以我们对每一个正样本需要生成多个负样本。生成的具体方式是将模块(a)步骤I的呈递表达数据对输入开源计算工具NetMHCpan(参考文献:Jurtz,Vanessa,et al."NetMHCpan-4.0:improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data."The Journal of Immunology 199.9(2017):3360-3368.),根据候选亲和力指数的倒序排名顺序,对每一个成功呈递的正样本数据对生成1-1000不等的负样本。但正负样本不平衡的数据对机器学习模型构建与优化是极大的挑战,故我们首先将正负样本比例设置为1:1,构建正负比相对平衡的数据集来训练预训练模型——这既是用于优化预训练模型的源域数据集。

[0065] 下一步,我们构建用于模拟真实环境与进行模型迁移的目标域数据集。相似的,我们首先将正负样本比例设置为1:100或1:10,以模拟真实生产环境里未呈递的负样本的出现概率远大于呈递成功的正样本的情况。需要声明的是,直接在此正负样本比例极端失衡的数据上,决大多数机器学习模型都难以稳定收敛,也难以学到有价值的正负样本中各特征的表达。

[0066] 另外,对于用于优化预训练模型的源域数据集以及进行模型迁移的目标域数据集,我们在候选负样本池中,选取特定数量负样本时,我们均可采取不同的采样生成策略。来模拟真实生产场景中的不同可能性。具体的,我们采取三种不同的负样本(阴性数据)策略:1.广义阴性策略:给定成功表达呈递的阳性数据对,该策略在其所有的对应阴性数据候选池中,随机选取一个作为数据集构建的阴性数据。2.中义阴性策略:给定成功表达呈递的阳性数据对,该策略在其所有的对应阴性数据候选池中,我们根据其亲和力指数(affinity score)进行降序排序,选择亲和力指数最小、与阳性数据相似度最低的阴性数据作为数据集构建。3.狭义阴性策略:给定成功表达呈递的阳性数据对,该策略在其所有的对应阴性数据候选中,我们首先剔除其原始亲和力指数(affinity score)<500的阴性数据样本,再对剩下的样本进行降序排序,选择亲和力指数最小、与阳性数据相似度最低的阴性数据作为数据集构建。

[0067] 为了进行后续“严格准确率(strict PPV)”指标的计算,我们还需额外构造一组特殊的测试数据集。具体方法是:从给定数据集中选取1000对成功呈递的蛋白质-HLA组队,对每一组组队再生成100个假阳性样本。

[0068] 具体的,我们选择MARIA,NetMHCpan(参考文献:Chen,Binbin,et al."Predicting HLA class II antigen presentation through integrated deep learning."Nature biotechnology 37.11(2019),Reynisson,Birkir,et al."NetMHCpan-4.1and NetMHCIIpan-4.0:improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data."Nucleic acids research 48.W1(2020):W449-W454.)数据集构建正负样本均衡的源域数据集,用于训练预训练模型;选择图宾根数据集(参考文献:Rammensee,H-G.,et al."SYFPEITHI:database for MHC ligands and peptide motifs."Immunogenetics 50.3

(1999):213-219.)。构建正负样本极端失衡的目标域数据集,以及strict PPV测试集。

[0069] 另外,对于基于图宾根数据集目标域数据集,我们选择k-折交叉验证(k-fold cross validatio)的统计学方法构建各自的训练、测试、验证数据集。其中训练、验证指用于步骤(b)中的模型迁移过程,测试指对步骤(c)中的预训练模型,迁移后模型进行测试。

[0070] (b)模型增强与迁移

[0071] 如图1所示,我们首先在源域数据集上优化预训练模型。

[0072] 具体的,对于源域数据集上的的不同模态特征,我们将其划分为:长序列特征,短序列特征,向量特征,标量特征,并对每种特征定义相应的归一化处理方案,以获取格式、维度统一,便于融合的特征向量。比如,对于给定多肽肽链氨基酸序列,使用随机矩阵将其每个氨基酸编码到可学习的到隐空间,再利用门循环单元网络(GRU)(参考文献:Chung,Junyoung,et al."Empirical evaluation of gated recurrent neural networks on sequence modeling."arXiv preprint arXiv:1412.3555(2014).)进行处理得到多肽序列特征。对于给定基因上下游肽链,利用独热方法(one-hot)进行编码,再将编码后的上下游序列进行裁剪得到定长的序列,再将此编码序列送入多层感知机网络模型进行变换,提取特征。对于亲和力指数特征(affinity score),考虑到原始数据尺度范围阔度较大:从几百到几万不等,我们采用$1-\log_{50}(kd)$,进行变换与尺度放缩。最后将处理好的各模态特征输入特征融合层,并最终得到如下优化模型:

[0073] $$\min_W \sum_{n=1}^{N} -w_n \left[y_n \cdot \log \sigma\left(f_W(x_n)\right) + (1-y_n) \cdot \log\left(1-\sigma\left(f_W(x_n)\right)\right)\right]$$

[0074] 其中f是步骤中集成了所有序列编码、多模态融合、特征变换神经网络的预测模型,W是该模型中可学习参数。w_n表示对不同样本的损失函数所赋予权重。在训练数据正负比均衡的情况下通常均赋值为1。在可能的训练数据正负比不均衡的情况下,可给赋值给较少的样本更大的权重。其中x_n是输入的特定数据(多肽、上下游、亲和力指数等),y_n是训练数据中是否呈递结合的真实值,o'是S逻辑函数(sigmoid function)。

[0075] 上述最优化模型的求解,可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 1hour."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。我们将优化后的模型参数与结构保存为预训练模型。

[0076] 注意,此步骤需要将源域数据集遍历多遍,反复优化,算力与时间成本较高。

[0077] 如图2所示,下一步我们在目标域数据集上对预训练模型进行深度迁移学习。

[0078] 对目标域数据集的多模态特征进行与源域数据集类似的数据处理与特征融合后,我们在其中随机抽样<5%比例的数据,进行深度迁移学习。我们可以根据预训练模型大小与数据规模,选择"全局优化"与"选定层优化"两种策略。前者优化预训练模型中所有可训练参数,后者仅优化神经网络模型中的最后两层,其他参数保持与预训练模型一致。整个过

程可视为优化以下模型：

[0079] $$\min_{W'} \sum_{i=1}^{N_1} \text{loss}_S \left(f'_{W'}, D_S^i, Y_S^i \right) + \lambda \sum_{i=1}^{N_2} \text{loss}_C \left(f'_{W'}, D_C^i, Y_C^i \right)$$

[0080] 式中，∫′是待迁移的预测模型；

[0081] 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；loss_S，loss_C分别表示在预训练阶段与模型迁移自适应阶段的目标损失函数(实验中两者皆取预训练阶段的优化模型)，其形式与上一步中介绍的一致；λ表示赋予模型迁移自适应阶段的目标损失函数的权重；$(D_S^i, Y_S^i), (D_C^i, Y_C^i)$分别表示构建的源域数据集与目标领域数据集上的训练数据特征与是否呈递结合的真实值；N_1，N_2分别表示构建的源域数据集与构建的目标领域数据集上的训练样本数量。通常前者远大于后者。在对模型进行优化后，将相关参数以结构化方法保存为迁移模型。

[0082] 上述最优化模型的求解，仍然采用批次随机梯度下降策略与ADMA优化器。但遍历目标域数据集的少量抽样数据的轮数小于5，算力与时间成本极低。

[0083] 为了进一步提升模型迁移效果，我们在实验中还会对源域数据集与目标领域数据集进行噪声白化(noise whitening)(参考文献：Alam，Md Jahangir，Gautam Bhattacharya，and Patrick Kenny."Speaker verification in mismatched conditions with frustratingly easy domain adaptation."Odyssey.Vol.2018.2018.)，数据流形对齐(manifold alignment)(参考文献：Wang，Chang，and Sridhar Mahadevan."Heterogeneous domain adaptation using manifold alignment."Twenty-second international joint conference on artificial intelligence.2011.)等领域自适应操作。

[0084] (c)模型检验

[0085] 我们直接在正负样本失衡的目标域数据集所划分的测试数据集上，用常规指标：接收者操作特征曲线(receiver operating characteristic curve，或者叫ROC曲线)下面积AUC与精准度PPV来评价迁移学习前后模型的预测能力与性能(表1)，具体做法是5次随机训练-测试划分别进行训练与测试，得到的5次结果取平均并计算标准差(在表格中显示为“平均值±标准差”)

[0086] 表1.预测模型评价指标

[0087]

评价指标	描述
精准度/PPV	TP/(TP+FP)
AUC	ROC曲线下面积

[0088] 以下是基于NetHMCPan数据源生成的正负比1:1的源域数据集，基于图宾根数据源生成的正负比1:10的目标域数据集兼测试集的实验结果(表2)：

[0089] 表2.NetHMCPan源域数据，TUBINGEN兼目标数据&测试集测试结果

[0090]

数据集生成策略/亲和力变换方法	预训练模型（未增强）	迁移模型（增强）
广义阴性策略	auc=0.735 ± 0.009 ppv=0.331 ± 0.008	auc=0.931 ± 0.009 ppv=0.869 ± 0.011
中义阴性策略	auc=0.841 ± 0.003 ppv=0.534 ± 0.012	auc=0.927 ± 0.005 ppv=0.780 ± 0.003
狭义阴性策略	auc=0.746 ± 0.009 ppv=0.297 ± 0.013	auc=0.922 ± 0.006 ppv=0.919 ± 0.011

[0091] 从表中可以看到，在所有阴性数据生成策略下，迁移学习前，基于1:1正负比训练数据的模型，在1:10正负比的测试数据上效果均很差。迁移学习后，在AUC、PPV上均有极大幅度的很大提升。说明了深度迁移学习作为我们应对正负比不平衡数据工具的有效性。由此证明了基于深度迁移的模型增强的有效性。

[0092] 本文中所描述的具体实施例仅仅是对本发明精神做举例说明。本发明所属技术领域的技术人员可以对所描述的具体实施例做各种各样的修改或补充或采用类似的方式替代，但并不会偏离本发明的精神或者超越所附权利要求书所定义的范围。

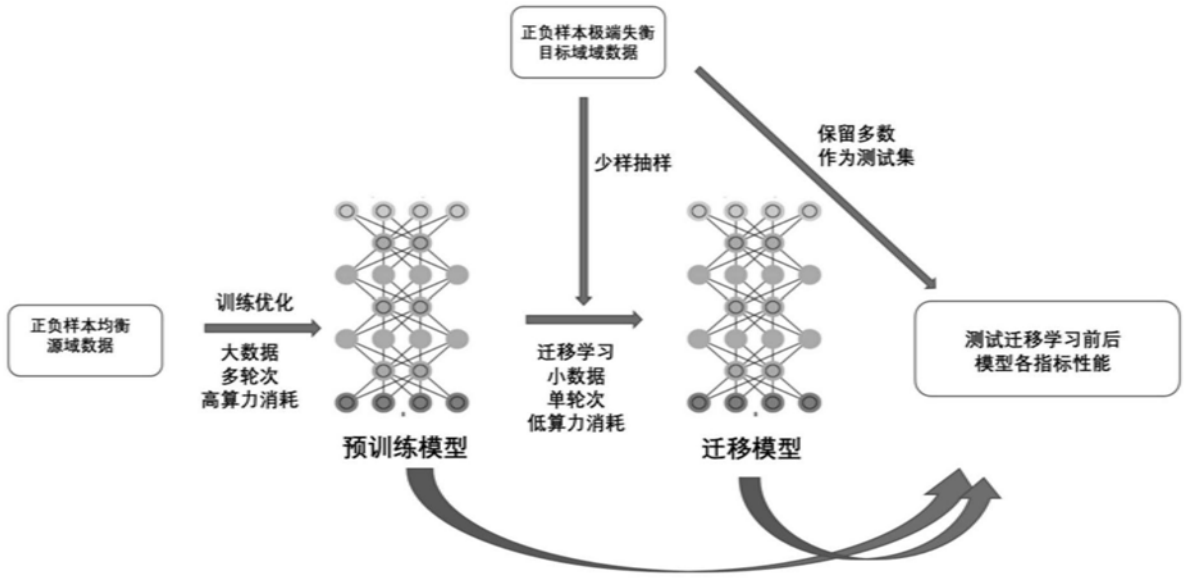

正负样本极端失衡
目标域域数据
少样抽样
保留多数
作为测试集
正负样本均衡
源域数据
训练优化
大数据
多轮次
高算力消耗
迁移学习
小数据
单轮次
低算力消耗
测试迁移学习前后
模型各指标性能
预训练模型
迁移模型

图1

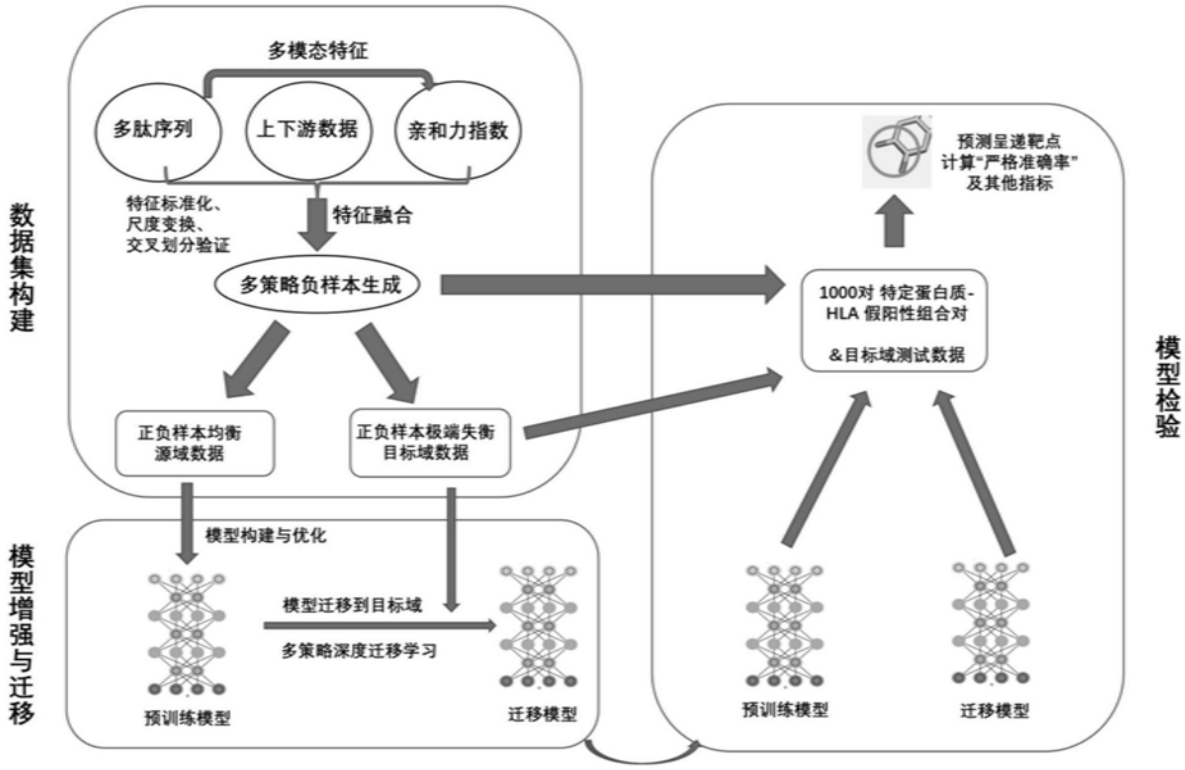

数据集构建
多模态特征
多肽序列
上下游数据
亲和力指数
特征标准化、
尺度变换、
交叉划分验证
特征融合
多策略负样本生成
正负样本均衡
源域数据
正负样本极端失衡
目标域数据
模型增强与迁移
模型构建与优化
模型迁移到目标域
多策略深度迁移学习
预训练模型
迁移模型
预测呈递靶点
计算"严格准确率"
及其他指标
1000对 特定蛋白质-
HLA 假阳性组合对
&目标域测试数据
预训练模型
迁移模型
模型检验

图2