

基于迁移学习的 MHC-I 型抗原表位呈递预测

胡伟澎^{1,2,3}, 李佑平^{2,3,4}, 张秀清^{2,3,4}

1. 华南理工大学生物科学与工程学院, 广州 510006
2. 深圳华大生命科学研究院, 深圳 518083
3. 华大吉诺因, 武汉 4300794
4. 中国科学院大学华大教育中心, 深圳 518083

摘要: 基于新抗原的肿瘤免疫治疗, 抗原呈递的准确预测是筛选 T 细胞特异性表位的关键步骤。质谱鉴定的表位数据对建立抗原呈递预测模型具有重要价值。尽管近年来质谱数据的积累持续增加, 但是大部分人类白细胞抗原(human leukocyte antigen, HLA)分型所对应的多肽数量相对较少, 无法建立可靠的预测模型。为此, 本研究尝试利用迁移学习的方法, 先利用混合分型的表位数据建立模型以识别抗原表位的共同特征, 在此预训练模型的基础上再利用分型特异性数据建立抗原呈递预测模型 Pluto。在相同的验证集上, Pluto 的平均 0.1% 阳性预测值(positive predictive value, PPV)比从头训练的模型高 0.078。在外部的质谱数据独立评估上, Pluto 的平均 0.1% PPV 为 0.4255, 高于从头训练模型(0.3824)和其他主流工具, 包括 MixMHCpred (0.3369)、NetMHCpan4.0-EL (0.4000)、NetMHCpan4.0-BA (0.3188)和 MHCflurry (0.3002)。此外, 在免疫原性预测评估上, Pluto 相对于其他工具也能找到更多的新抗原。Pluto 开源网址: <https://github.com/weipengHU/Pluto>。

关键词: 免疫治疗; 新抗原; 抗原呈递; 深度学习; 迁移学习

MHC-I epitope presentation prediction based on transfer learning

Weipeng Hu^{1,2,3}, Youping Li^{2,3,4}, Xiuqing Zhang^{2,3,4}

1. School of Biology and Biological Engineering, South China University of Technology, Guangzhou, 510006, China
2. BGI-Shenzhen, Shenzhen 518083, China
3. BGI-GenoImmune, Wuhan 4300794, China
4. BGI Education Center, University of Chinese Academy of Sciences, Shenzhen 518083, China

Abstract: Accurate epitope presentation prediction is a key procedure in tumour immunotherapies based on neoantigen

收稿日期: 2019-06-21; 修回日期: 2019-09-17

基金项目: 国家自然科学基金项目(编号: 81702826, 81772910), 深圳市科创委项目(编号: JCYJ20170303151334808)和深圳市经信委项目(编号: 20170731162715261)资助[Supported by the National Natural Science Foundation of China (Nos. 81702826, 81772910), Science, Technology and Innovation Commission of Shenzhen Municipality (No. JCYJ20170303151334808) and Shenzhen Municipal Government of China (No. 20170731162715261)]

作者简介: 胡伟澎, 硕士研究生, 专业方向: 基因组学。E-mail: huweipeng@genomics.cn

李佑平, 硕士研究生, 专业方向: 基因组学。E-mail: liyouping@genomics.cn

胡伟澎和李佑平并列第一作者。

通讯作者: 张秀清, 博士, 教授, 研究方向: 基因组学及免疫治疗。E-mail: zhangxq@genomics.cn

DOI: 10.16288/j.ycz.19-155

网络出版时间: 2019/11/8 13:27:56

URI: <http://kns.cnki.net/kcms/detail/11.1913.R.20191107.1628.005.html>

for targeting T cell specific epitopes. Epitopes identified by mass spectrometry (MS) is valuable to train an epitope presentation prediction model. In spite of the accelerating accumulation of MS data, the number of epitopes that match most of human leukocyte antigens (HLAs) is relatively small, which makes it difficult to build a reliable prediction model. Therefore, this research attempted to use the transfer learning method to train a model to learn common features among the mixed allele specific epitopes. Then based on this pre-trained model, we used the allele-specific epitopes to train the final epitope presentation prediction model, termed Pluto. The average 0.1% positive predictive value (PPV) of Pluto outperformed the prediction model without pretraining with a margin of 0.078 on the same validation dataset. When evaluating Pluto on external HLA eluted ligand datasets, Pluto achieved an averaged 0.1% PPV of 0.4255, which is better than the prediction model without pretraining (0.3824) and other popular methods, including MixMHCpred (0.3369), NetMHCpan4.0-EL (0.4000), NetMHCpan4.0-BA (0.3188) and MHCflurry (0.3002). Moreover, when it comes to the evaluation of predicting immunogenicity, Pluto can identify more neoantigens than other tools. Pluto is publicly available at <https://github.com/weipenegHU/Pluto>.

Keywords: immunotherapy; neoantigen; epitope presentation; deep learning; transfer learning

肿瘤细胞内含有肿瘤特异性突变位点的蛋白质能够被消化为不同长度的多肽, 含有突变的多肽能够在内质网中与主要组织相容性复合体(major histocompatibility complex, MHC)结合形成多肽-MHC复合物然后被呈递到细胞表面, 如果多肽-MHC复合物被 T 细胞特异性识别, 即能够引起肿瘤细胞的凋亡, 这种多肽被称为新抗原。近年来, 基于新抗原的肿瘤免疫疗法在不同癌种的治疗中取得令人瞩目的突破^[1-7], 而且新抗原对于预测肿瘤疗效和病人预后具有重要价值^[8-12]。目前筛选新抗原的主流方法是通过亲和力预测工具预测多肽能否和人类白细胞抗原(human leukocyte antigen, HLA)结合, 例如 NetMHCpan 系列工具^[13-15]和 MHCflurry^[16]等。但是, 这些工具使用的训练数据大部分来源于体外实验, 不能真实反应细胞内多肽与 HLA 结合的情况。随着质谱技术的发展, 科学家们能够直接获得呈递到细胞表面的多肽数据, 相对于传统的经体外实验得到的亲和力数据, 这些质谱数据更加真实地反应了多肽在细胞内加工到呈递的自然过程, 包含更多的信息。随着质谱数据的积累以及质谱数据对多肽免疫原性预测的重要性得到越来越多的重视^[17,18], 基于质谱数据训练的抗原呈递预测模型也随之出现, 例如 MixMHCpred^[19,20]和 EDGE^[17]。

虽然目前已经积累了一定数量的质谱鉴定的抗原表位数据, 但对应到每个 HLA 分型的质谱数据并

不均匀, 大部分的 HLA 分型只有数千条多肽数据, 有的更只有数百条。在这种情况下, 并不能开发出可靠的分型特异性的抗原呈递预测模型。迁移学习或许能够帮助改善目前的这种状况, 其基本原理是利用在一个相似任务上学习到的经验转移到最终需要解决的任务上, 通常前者拥有大量的数据, 而后者只有少量的数据。为了验证上述猜想, 本研究先利用混合分型的 MHC-I 亚型抗原表位数据(是指训练数据由对应不同 MHC-I 亚型的抗原表位组成)来训练一个模型以区分抗原表位与普通的蛋白质多肽, 再利用另外的包括 16 个 HLA 分型的单分型抗原表位数据在预训练模型的基础上训练最终的分型特异性抗原呈递模型, 称之为 Pluto。然后, 在相同的验证集上评估了 Pluto 相对于从头训练模型的优势, 并在独立验证集上比较了其与目前主流软件的表现。Pluto 模型有望为相关工作提供新的思路以及对免疫治疗领域做出有益的贡献。

1 材料与方法

1.1 训练集构建

预训练模型用到的阳性集来源于 Pearson 等^[21]和 Bassani-Sternberg 等^[22]产生的数据以及 SystemMHC 质谱多肽数据库^[23]。将这些数据集合并后, 剔除长

度小于 8 以及大于 14 的多肽, 然后根据多肽和 HLA 分型去重, 总共得到接近 16 万的多肽数据(表 1)。阴性集来源于人类蛋白组的随机切割的多肽(剔除出现在阳性数据集中的多肽), 从中挑取与阳性集等量的阴性多肽与阳性集合并构成训练集, 然后从训练集中各挑取 5000 条阳性多肽和 5000 条阴性多肽构成验证集。

抗原呈递模型中用到的阳性训练数据来源于 Abelin 等^[24]研究的 16 个单分型细胞系, 包括 A01:01、A02:01、A02:03、A02:04、A02:07、A03:01、A24:02、A29:02、A31:01、A68:02、B35:01、B44:02、B44:03、B51:01、B54:01 和 B57:01, 总共约有 2.7 万条的多肽(表 1), 分别为这 16 个分型单独建模。每个分型的数据按照 8:2 的比例划分为阳性训练集和阳性验证集。从随机切割的蛋白质多肽中挑取阳性训练集数据 100 倍的阴性多肽与阳性训练集合并构成训练集, 挑取阳性验证集数据 999 倍的阴性多肽与阳性验证集合并构成验证集。

本研究构建的模型主要对长度在 8~14 的短肽进行预测, 因此预训练模型和抗原呈递模型使用的多肽先利用通配符‘X’把多肽的长度统一为 14 肽, 然后利用热编码将每条多肽编码为 294 维(14×21, 算上通配符‘X’, 每个氨基酸需要编码为 21 维向量)的向量。

1.2 模型训练

预训练的模型由输入层、5 层隐藏层和输出层组成(图 1A), 其中 5 层隐藏层包含的神经元数目分别为 100、30、100、30 和 10, 第一个和第三个隐藏层采用 dropout(dropout rate=0.4)来控制模型的过拟合, 各隐藏层均使用 exponential linear unit (ELU) 作为激活函数。本研究采用批次梯度下降的方法训练预训练模型, 每个批次包含 1024 条多肽(阳性和

阴性多肽各一半), 总共迭代 100 次。采用 5 层交叉验证的方法来评估预训练模型的准确率。

Pluto 的结构是在预训练模型结构的基础上, 在最后一层隐藏层和输出层之间增加了一个隐藏层, 这层隐藏层之前的神经元参数均使用预训练模型中对应神经元的参数, 而且不再对这些神经元进行训练, 而只对新增的隐藏层和输出层的神经元训练。同样采用批次梯度下降的方法训练模型, 每个批次包含全部的阳性多肽以及 10 倍的阴性多肽, 总共迭代 1000 次。每迭代一次, 利用训练好的模型对验证集进行预测打分, 根据分数大小进行排序并统计排名前 0.1% 的结果(阳性多肽的数目)中的阳性预测值(positive predictive value, PPV)。最后根据 0.1% PPV 的表现选择最终的模型。采用 0.1% PPV 评估标准是因为根据之前报道^[25-27], 细胞内能被呈递的多肽若占整个人类蛋白组的 0.1%, 因此该评估标准更能够

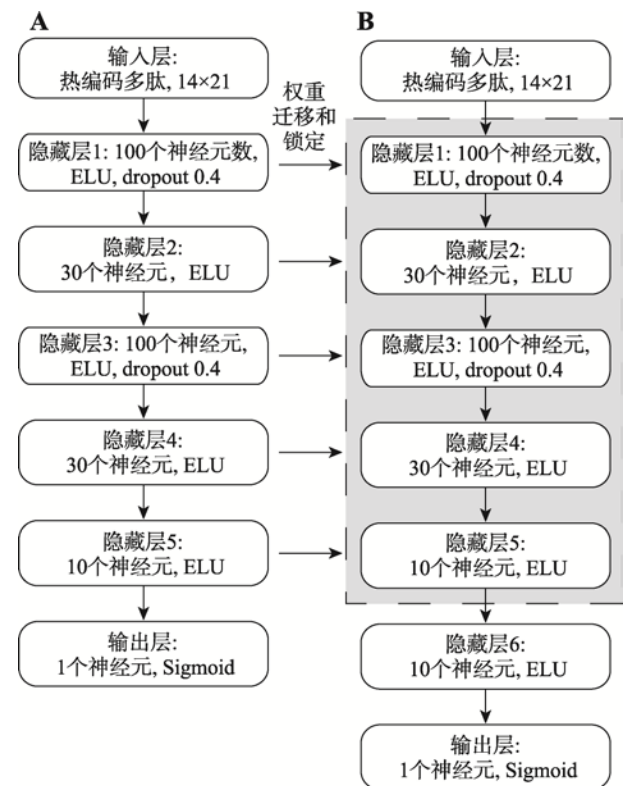


图 1 Pluto 的构建过程

Fig. 1 The construction of Pluto

A: 预训练模型的结构; B: Pluto 的结构中, 前 5 层隐藏层使用的权重来源于预训练模型的结构, 训练过程中对这些迁移过来的权重锁定, 即这些权重在训练过程中不会改变, 并且只对新增加的隐藏层和输出层的权重进行训练。

表 1 训练集总结

Table 1 Summary of training data set

数据来源	HLA 分型数	多肽数量	文献
Pearson 等	27	7833	[21]
Bassani-sternberg 等	4	49 711	[21]
SysteMHC	67	101 579	[21]
Abelin 等	16	26 661	[21]

反映实际情况。模型的实现和训练均采用 Tensorflow 框架^[28]。

1.3 外部质谱数据评估

收集了 Trolle 等^[29]产生的 HeLa 单分型细胞系多肽数据对模型进行独立评估。把这些质谱多肽与 999 倍的阴性多肽合并构建测试集,生成的测试集用于评估 Pluto、从头训练模型以及 MixMHCpred、NetMHCpan4.0-EL、NetMHCpan4.0-BA 和 MHCflurry 的 0.1% PPV。

1.4 免疫原性评估

收集了 Stronen 等^[30]和 Gros 等^[31]经实验验证具有免疫原性的多肽。因为 Stronen 等是利用四聚体实验直接对包含突变的多肽进行验证的,所以每条多肽是否具有免疫原性是明确的。而 Gros 等是利用串联迷你基因(tandem mini-gene, TMG)验证的,把这些 TMG 切成长度为 8~11 个氨基酸,包含突变位点的重叠连续多肽。来自于没有免疫原性的 TMG 的多肽被标记为阴性数据。来自于具有免疫原性的 TMG 但是没有经过多肽负载实验验证的多肽会被剔除,因为不能确定这些多肽能否被 T 细胞识别,其他多肽则按照多肽负载实验验证的结果标记为阳性和阴性多肽。然后利用 Pluto、MixMHCpred、NetMHCpan4.0-EL、NetMHCpan4.0-BA 和 MHCflurry 对这些多肽进行预测,并比较这些工具对免疫原性多肽的排位。

2 结果与分析

2.1 预训练能提高抗原呈递预测模型的表现

假设预训练模型从大量呈递的抗原表位中学习一些非分型特异性的特征,并且能够提高分型特异性抗原呈递预测模型的表现。为验证此假设,本研究利用单分型训练集从头训练 Pluto 整个网络的全部参数,而不利用预训练模型训练好的参数,并且在相同的验证集上和 Pluto 的表现作比较(图 2A)。

通过分析,经过混合分型的表位数据训练的预训练模型五层交叉验证的平均准确率为 90.77%,说明模型学习到一些能够将抗原表位与普通蛋白质多

肽区分开来的特征。接下来在 16 个单分型验证集上评估 Pluto 与从头训练模型的 0.1% PPV,结果发现 Pluto 的 0.1% PPV 在所有验证集上都比没有经预训练的模型要高,平均 0.1% PPV 提升了 0.078。然后观察了训练集大小与模型表现提升之间的关系,从图 2B 中可以发现这样一种趋势:迁移学习对数据量小的分型的表现提升帮助更加明显,而对数据量较大的分型来说,迁移学习对模型的提升则比较小。

因此,上述结果表明预训练模型能够学习到不同分型抗原表位的共同特征,并且能够帮助提高分型特异性的抗原呈递预测模型的表现,而提升的幅度可能受到抗原呈递预测模型的训练集大小的影响。

2.2 在质谱数据上独立评估模型表现

利用 Trolle 等^[29]产生的单分型质谱数据对 Pluto 的使用效果进行评估,并与从头训练的模型和主流预测工具作(包括 MixMHCpred (v2.0)^[19,20]、NetMHCpan4.0-EL^[13]、NetMHCpan4.0-BA^[13]和 MHCflurry^[16])进行比较,结果发现 Pluto 在独立测试集上的平均 0.1% PPV 为 0.4255,显著优于从头训练模型、MixMHCpred、NetMHCpan4.0-BA 和 MHCflurry,这些模型的平均 0.1% PPV 分别为 0.3824、0.3369、0.3188、0.3002 ($P = 0.02538$ 、 0.002035 、 0.01102 、 0.01929 , paired t -test)。值得注意的是,虽然 Pluto 的平均 0.1% PPV 没有显著高于 NetMHCpan4.0-EL (0.4255 vs. 0.4000, $P = 0.05311$),但是在每个分型上 Pluto 的表现都要好于 NetMHCpan4.0-EL (图 3)。

MixMHCpred 是基于位置特异性打分矩阵(position specific scoring matrix, PSSM)以及只用质谱数据训练的抗原呈递预测模型。PSSM 属于线性模型的一种,它基于的假设是多肽的每个位置都是独立,而从图 3 的分析结果看,从头训练模型和 NetMHCpan4.0-EL 表现要显著优于 MixMHCpred ($P = 0.03048$, $5.674e-05$, paired t -test),因此推测多肽的不同位置之间可能存在一定的联系,而不是单纯的线性关系(本研究选择从头训练模型与 NetMHCpan4.0-EL 和 MixMHCpred 比较,是因为它们都是抗原呈递预测模型,而 NetMHCpan4.0-BA 和 MHCflurry 是亲和力预测模型)。

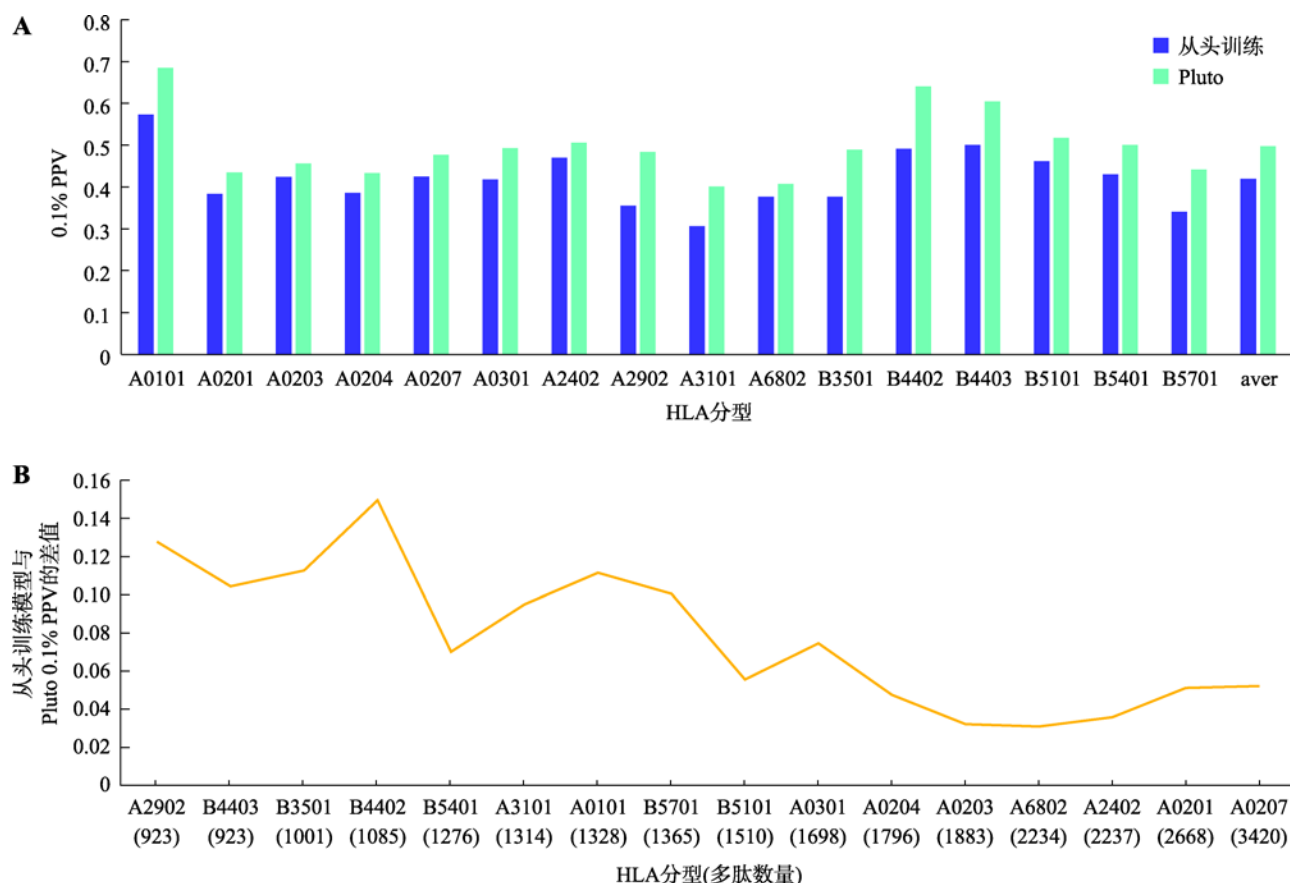


图 2 Pluto 与从头训练模型的性能比较

Fig. 2 Performance comparison between Pluto and model without pretraining

A: 在 16 个单分型相同的验证集上 Pluto 的 0.1% PPV 表现都要优于从头训练的模型; B: 预训练模型对 Pluto 表现提升的幅度受训练集大小的影响。

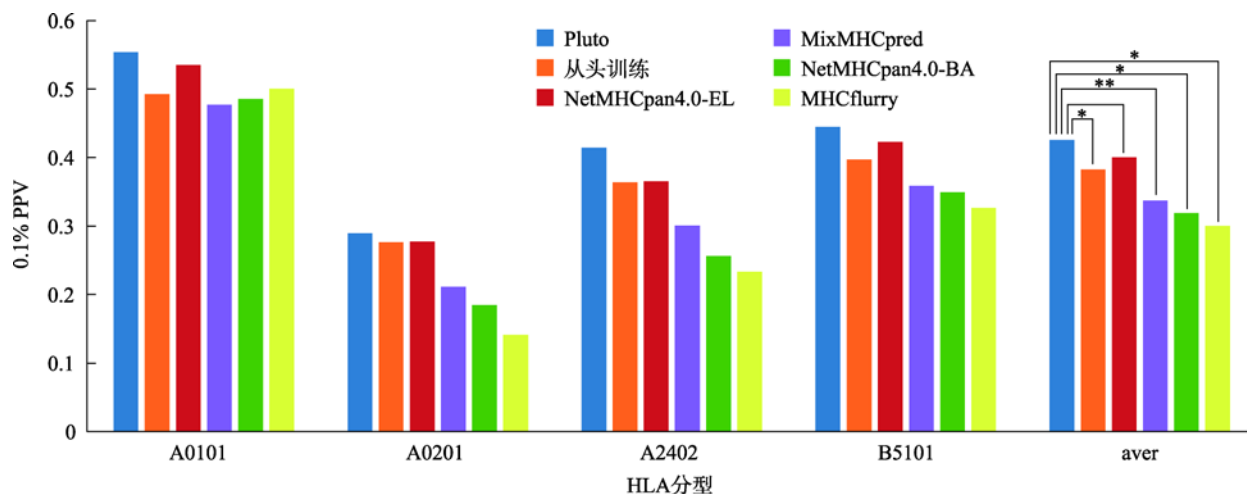


图 3 在外部质谱数据上进行独立评估

Fig. 3 Independent evaluation on external mass spectrometry data set

Pluto 的平均 0.1% PPV 要显著高于从头训练模型, MixMHCpred, NetMHCpan4.0-BA 和 MHCflurry。Pluto 的平均 0.1% PPV 虽然没有显著高于 NetMHCpan4.0-EL, 但是在每个分型上的表现都要高于 NetMHCpan4.0-EL。*代表 $P < 0.05$, **代表 $P < 0.005$ (paired t -test)。

综上所述,通过独立评估,本研究验证了 Pluto 能够达到甚至优于目前主流的抗原呈递预测工具的水平。

2.3 肿瘤新抗原鉴定

为评估 Pluto 预测抗原呈递的能力能否用于寻找新抗原,本研究从 Stronen 等^[30]和 Gros 等^[31]的研究中收集了 7 条经实验验证具有免疫原性的多肽,并利用这些多肽评估 Pluto、MixMHCpred、NetMHCpan4.0-EL、NetMHCpan4.0-BA 和 MHCflurry 预测新抗原的能力。结果如表 2 所示,在每个病人排名前 10 的多肽中,Pluto 能够找回 7 条免疫原性多肽中的 4 条,MixMHCpred 和 NetMHCpan4.0-EL 能够找回其中的两条,而 NetMHCpan4.0-BA 和 MHCflurry 只能找到其中的 1 条。因此,评估结果证明了 Pluto 对于鉴定肿瘤新抗原具有重要价值。

3 讨论

抗原呈递的准确预测是判断新抗原能否激活新抗原特异性 T 细胞从而杀死肿瘤细胞的关键一步。虽然近几年来质谱技术飞速发展,积累了不少通过质谱鉴定的抗原表位数据,但是对于特定分型来说,每个分型对应的抗原表位数据还不是很多,对于建立一个基于深度学习的分型特异性的抗原表位预测模型来说是不足够的。因此本研究利用迁移学习的方法,从大量的混合分型抗原表位数据和蛋白质组中随机多肽数据建立了一个深度学习模型以识别抗原表位是否存在一些共性,使之能够与普通的多肽区分开。然后在预训练模型的基础上,利用分型特

异性的数据训练了抗原呈递预测模型 Pluto。

本研究首先展示了预训练模型能够将大部分的抗原表位与蛋白质组的普通多肽分开,说明模型学到了抗原表位的一些共同特征。但是因为深度学习本身的原因,预训练模型学习到哪些共同特征尚无法明确,值得后续研究给予重点关注。Pluto 的表现相对于从头训练的模型的表现有明显的提升,但是提升的幅度受到分型特异性的训练集大小的影响。分析造成这种影响的原因可能有 3 个:一是随着分型特异性的抗原表位数据增加,所包含的信息量更多,与混合分型的抗原表位提供的信息有更大重合,这导致预训练模型学习到的特征起到的作用更小;二是模型可能已经接近饱和状态,增加数据量对模型提高帮助不大;三是随着数据量的增加,需要建立更加复杂的网络以学习更多的特征才能提高模型的表现。在利用外部数据进行独立评估以及鉴定新抗原上,Pluto 的表现也优于从头训练的模型以及这个领域的其他主流工具。本文中用到的所有训练数据和评估数据都可以从 <https://github.com/weipen-egHU/Pluto> 获取。

抗原表位需要经过源蛋白的表达,源蛋白经蛋白酶体消化切割后产生的多肽被转运到内质网内部与 MHC-I 分子结合,最后才能被呈递到细胞表面。在本研究中,Pluto 只是根据抗原表位序列自身包含的信息来判定多肽能否被呈递到细胞表面,而序列本身提供的信息是非常有限的。据文献报道,多肽的表达量对抗原呈递具有很大的影响^[17,24,32]。此外,抗原表位的上下游序列能够帮助预测多肽能否被蛋白酶体切割^[17,24,33]。还有文献报道能够产生抗原表位的蛋白质只占细胞内所有蛋白质的一部分^[21],以

表 2 Pluto 与主流工具对免疫原性多肽的排名
Table 2 Immunogenic peptides ranked by Pluto and other popular methods

病人编号	多肽	HLA	Pluto	MixMHCpred	NetMHCpan4.0 EL	NetMHCpan4.0 BA	MHCflurry
3903	AYHSIEWAI	HLA-A2402	76	32	22	147	113
3998	KVDPIGHVY	HLA-A0101	3	2	1	1	4
3998	KVDPIGHVYIF	HLA-A0101	5	148	27	31	113
patient1	ALDPHSGHFV	HLA-A0201	9	17	14	25	17
patient1	ALLETPSLL	HLA-A0201	18	5	6	10	22
patient1	ALLETPSLLL	HLA-A0201	3	20	12	16	29
patient2	ELMRDINSM	HLA-B3501	128	107	57	30	45

及蛋白质中存在产生抗原表位的热点^[34]。相信这些特征能够进一步提高 Pluto 的表现, 开发和利用这些特征将是未来工作的重要方向。

虽然根据抗原呈递的可能性挑选新抗原具有一定效果^[17], 但是被呈递的多肽不一定具有免疫原性(多肽的免疫原性是指多肽能否被 T 细胞识别从而杀死肿瘤细胞)^[35-37]。所以除了抗原呈递预测外, 对多肽的免疫原性预测也具有重要意义。但是目前因为免疫原性数据缺乏积累, 所以难以建立多肽免疫原性预测模型。未来通过共同协作产生更多的免疫原性数据, 更好的实验方法来了解 TCR 和多肽-MHC 分子的相互作用^[38,39]以产生更大的数据集和对免疫原性更深的生物学认识, 最终能够更准确地预测免疫原性。

综上所述, 本研究利用迁移学习的方法建立了一个新的抗原呈递预测工具 Pluto, 其表现显著优于目前主流的预测软件。同时, 这些结果说明了迁移学习对解决目前因分型特异性的抗原表位数据不足而难以建立一个可靠的抗原呈递预测模型的问题有所帮助。

参考文献(References):

- [1] Gros A, Parkhurst MR, Tran E, Pasetto A, Robbins PF, Ilyas S, Prickett TD, Gartner JJ, Crystal JS, Roberts IM, Trebska-Mcgowan K, Wunderlich JR, Yang JC, Rosenberg SA. Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. *Nat Med*, 2016, 22(4): 433-438. [DOI]
- [2] Malekzadeh P, Pasetto A, Robbins PF, Parkhurst MR, Paria BC, Jia L, Gartner JJ, Hill V, Yu Z, Restifo NP, Sachs A, Tran E, Lo W, Somerville RPT, Rosenberg SA, Deniger DC. Neoantigen screening identifies broad TP53 mutant immunogenicity in patients with epithelial cancers. *J Clin Invest*, 2019, 129(3): 1109-1114. [DOI]
- [3] Robbins PF, Lu YC, El-Gamil M, Li YF, Gross C, Gartner J, Lin JC, Teer JK, Clifton P, Tycksen E, Samuels Y, Rosenberg SA. Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat Med*, 2013, 19(6): 747-752. [DOI]
- [4] Sahin U, Derhovanessian E, Miller M, Kloke BP, Simon P, Löwer M, Bukur V, Tadmor AD, Luxemburger U, Schrörs B, Omokoko T, Vormehr M, Albrecht C, Paruzynski A, Kuhn AN, Buck J, Heesch S, Schreeb KH, Müller F, Ortseifer I, Vogler I, Godehardt E, Attig S, Rae R, Breitkreuz A, Tolliver C, Suchan M, Martic G, Hohberger A, Sorn P, Diekmann J, Ciesla J, Waksman O, Brück A K, Witt M, Zillgen M, Rothermel A, Kasemann B, Langer D, Bolte S, Diken M, Kreiter S, Nemecek R, Gebhardt C, Grabbe S, Höller C, Utikal J, Huber C, Loquai C, Türeci O. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*, 2017, 547(7662): 222-226. [DOI]
- [5] Tran E, Ahmadzadeh M, Lu YC, Gros A, Turcotte S, Robbins PF, Gartner JJ, Zheng Z, Li YF, Ray S, Wunderlich JR, Somerville RP, Rosenberg SA. Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science*, 2015, 350(6266): 1387-1390. [DOI]
- [6] Tran E, Robbins PF, Lu YC, Prickett TD, Gartner JJ, Jia L, Pasetto A, Zheng Z, Ray S, Groh EM, Kriley IR, Rosenberg SA. T-Cell transfer therapy targeting mutant KRAS in cancer. *N Engl J Med*, 2016, 375(23): 2255-2262. [DOI]
- [7] Zacharakis N, Chinnasamy H, Black M, Xu H, Lu YC, Zheng Z, Pasetto A, Langhan M, Shelton T, Prickett T, Gartner J, Jia L, Trebska-Mcgowan K, Somerville RP, Robbins PF, Rosenberg SA, Goff SL, Feldman SA. Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer. *Nat Med*, 2018, 24(6): 724-730. [DOI]
- [8] Strickland KC, Howitt BE, Shukla SA, Rodig S, Ritterhouse LL, Liu JF, Garber JE, Chowdhury D, Wu CJ, D'andrea AD. Association and prognostic significance of BRCA1/2-mutation status with neoantigen load, number of tumor-infiltrating lymphocytes and expression of PD-1/PD-L1 in high grade serous ovarian cancer. *Oncotarget*, 2016, 7(12): 13587-13598. [DOI]
- [9] Lu HZ, Wang DK, Wang Z. Correlation analysis of the prognosis of HPV positive oropharyngeal cancer patients with T cell infiltration and neoantigen load. *Hereditas (Beijing)*, 2019, 41(8): 725-735. 卢涣滋, 王迪侃, 王智. HPV 阳性口咽癌患者预后与 T 细胞浸润和新抗原负荷相关性分析. *遗传*, 2019, 41(8): 725-735. [DOI]
- [10] Brown SD, Warren RL, Gibb EA, Martin SD, Spinelli JJ, Nelson BH, Holt RA. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res*, 2014, 24(5): 743-750. [DOI]
- [11] Shukla SA, Howitt BE, Wu CJ, Konstantinopoulos PA.

- Predicted neoantigen load in non-hypermutated endometrial cancers: Correlation with outcome and tumor-specific genomic alterations. *Gynecol Oncol Rep*, 2016, 19: 42–45. [DOI]
- [12] Sa HL, Ma KW, Gao Y, Wang DQ. Predictive value of tumor mutation burden in immunotherapy for lung cancer. *Chin J Lung Canc*, 2019, 22(6): 380–384. 撒焕兰, 马克威, 高勇, 王德强. 肿瘤突变负荷对肺癌免疫治疗疗效的预测价值. *中国肺癌杂志*, 2019, 22(6): 380–384. [DOI]
- [13] Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol*, 2017, 199(9): 3360–3368. [DOI]
- [14] Nielsen M, Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med*, 2016, 8(1): 33. [DOI]
- [15] Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, Røder G, Peters B, Sette A, Lund O, Buus S. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One*, 2007, 2(8): e796. [DOI]
- [16] O'donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst*, 2018, 7(1): 129–132 e4. [DOI]
- [17] Bulik-Sullivan B, Busby J, Palmer CD, Davis MJ, Murphy T, Clark A, Busby M, Duke F, Yang A, Young L, Ojo NC, Caldwell K, Abhyankar J, Boucher T, Hart MG, Makarov V, Montpreville VT, Mercier O, Chan TA, Scagliotti G, Bironzo P, Novello S, Karachaliou N, Rosell R, Anderson I, Gabrail N, Hrom J, Limvarapuss C, Choquette K, Spira A, Rousseau R, Voong C, Rizvi NA, Fadel E, Frattini M, Jooss K, Skoberne M, Francis J, Yelensky R. Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat Biotechnol*, 2018, 37(1): 55–63. [DOI]
- [18] Gfeller D, Bassani-Sternberg M. Predicting antigen presentation—what could we learn from a million peptides? *Front Immunol*, 2018, 9: 1716. [DOI]
- [19] Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak H, Gannon PO, Kandalaft LE, Coukos G, Gfeller D. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allosteric regulating HLA specificity. *PLoS Comput Biol*, 2017, 13(8): e1005725. [DOI]
- [20] Gfeller D, Guillaume P, Michaux J, Pak HS, Daniel RT, Racle J, Coukos G and Bassani-Sternberg M. The length distribution and multiple specificity of naturally presented HLA-I ligands. *J Immunol*, 2018, 201(12): 3705–3716. [DOI]
- [21] Pearson H, Daouda T, Granados DP, Durette C, Bonneil E, Courcelles M, Rodenbrock A, Laverdure JP, Coté C, Mader S, Lemieux S, Thibault P, Perreault C. MHC class I-associated peptides derive from selective regions of the human genome. *J Clin Invest*, 2016, 126(12): 4690–4701. [DOI]
- [22] 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, 14(3): 658–673. [DOI]
- [23] Shao W, Pedrioli PGA, Wolski W, Scurtescu C, Schmid E, Vizcaíno JA, Courcelles M, Schuster H, Kowalewski D, Marino F, Arlehamn CSL, Vaughan K, Peters B, Sette A, Ottenhoff THM, Meijgaarden KE, Nieuwenhuizen N, Kaufmann SHE, Schlapbach R, Castle JC, Nesvizhskii A I, Nielsen M, Deutsch E W, Campbell D S, Moritz R L, Zubarev R A, Ytterberg A J, Purcell A W, Marcilla M, Paradela A, Wang Q, Costello CE, Ternette N, van Veelen PA, van Els CACM, Heck AJR, de Souza GA, Sollid LM, Admon A, Stevanovic S, Rammensee HG, Thibault P, Perreault C, Bassani-Sternberg M, Aebersold R, Caron E. The SystemMHC atlas project. *Nucleic Acids Res*, 2018, 46(D1): D1237–D1247. [DOI]
- [24] Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, Stevens J, Lane W, Zhang GL, Eisenhaure TM, Clauser KR, Hacohen N, Rooney MS, Carr SA, Wu CJ. Mass spectrometry profiling of HLA-Associated peptidomes in Mono-allelic cells enables more accurate epitope prediction. *Immunity*, 2017, 46(2): 315–326. [DOI]
- [25] Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, Wheeler DK, Gabbard JL, Hix D, Sette A, Peters B. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res*, 2015, 43(Database issue): D405–412. [DOI]
- [26] Rammensee HG, Friede T, Stevanović S. MHC ligands and peptide motifs: first listing. *Immunogenetics*, 1995, 41(4): 178–228. [DOI]
- [27] 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(5049): 1261–1263. [DOI]
- [28] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia YQ, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mane D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viegas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng XQ. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. [DOI]
- [29] Trolle T, McMurtrey CP, Sidney J, Bardet W, Osborn SC, Kaeffer T, Sette A, Hildebrand WH, Nielsen M, Peters B. The length distribution of class I-restricted T cell epitopes is determined by both peptide supply and MHC allele-specific binding preference. *J Immunol*, 2016, 196(4): 1480–1487. [DOI]
- [30] Strønen E, Toebes M, Kelderman S, van Buuren MM, Yang W, van Rooij N, Donia M, Bösch ML, Lund-Johansen F, Olweus J, Schumacher TN. Targeting of cancer neoantigens with donor-derived T cell receptor repertoires. *Science*, 2016, 352(6291): 1337–1341. [DOI]
- [31] Gros A, Parkhurst MR, Tran E, Pasetto A, Robbins PF, Ilyas S, Prickett TD, Gartner JJ, Crystal JS, Roberts IM. Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. *Nat Med*, 2016, 22(4): 433–438. [DOI]
- [32] Hu WP, Qiu S, Li YP, Lin XX, Zhang L, Xiang HT, Han X, Chen L, Li S, Li WH, Ren Z, Hou GX, Lin ZL, Lu JL, Liu G, Li B, Lee LJ. EPIC: MHC-I epitope prediction integrating mass spectrometry derived motifs and tissue-specific expression profiles. *bioRxiv*, 2019, 567081. [DOI]
- [33] Nielsen M, Lundegaard C, Lund O, Kesmir C. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics*, 2005, 57(1–2): 33–41. [DOI]
- [34] Müller M, Gfeller D, Coukos G, Bassani-Sternberg M. 'Hotspots' of antigen presentation revealed by human leukocyte antigen ligandomics for neoantigen prioritization. *Front Immunol*, 2017, 8: 1367. [DOI]
- [35] Mcgranahan N, Furness AJ, Rosenthal R, Ramskov S, Lyngaa R, Saini SK, Jamal-Hanjani M, Wilson GA, Birkbak NJ, Hiley CT, Watkins TB, Shafi S, Murugaesu N, Mitter R, Akarca AU, Linares J, Marafioti T, Henry JY, Van Allen EM, Miao D, Schilling B, Schadendorf D, Garraway LA, Makarov V, Rizvi NA, Snyder A, Hellmann MD, Merghoub T, Wolchok JD, Shukla SA, Wu CJ, Peggs KS, Chan TA, Hadrup SR, Quezada SA, Swanton C. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*, 2016, 351(6280): 1463–1469. [DOI]
- [36] Calis JJ, Maybeno M, Greenbaum JA, Weiskopf D, de Silva AD, Sette A, Keşmir C, Peters B. Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput Biol*, 2013, 9(10): e1003266. [DOI]
- [37] Assarsson E, Sidney J, Oseroff C, Pasquetto V, Bui HH, Frahm N, Brander C, Peters B, Grey H, Sette A. A quantitative analysis of the variables affecting the repertoire of T cell specificities recognized after vaccinia virus infection. *J Immunol*, 2007, 178(12): 7890–7901. [DOI]
- [38] Bentzen AK, Such L, Jensen KK, Marquard AM, Jessen LE, Miller NJ, Church CD, Lyngaa R, Koelle DM, Becker JC, Linnemann C, Schumacher TNM, Marcanti P, Nghiem P, Nielsen M, Hadrup SR. T cell receptor fingerprinting enables in-depth characterization of the interactions governing recognition of peptide–MHC complexes. *Nat Biotechnol*, 2018, 36(12): 1191–11996. [DOI]
- [39] Bentzen AK, Marquard AM, Lyngaa R, Saini SK, Ramskov S, Donia M, Such L, Furness AJ, Mcgranahan N, Rosenthal R, Straten PT, Szallasi Z, Svane IM, Swanton C, Quezada SA, Jakobsen SN, Eklund AC, Hadrup SR. Large-scale detection of antigen-specific T cells using peptide–MHC-I multimers labeled with DNA barcodes. *Nat Biotechnol*, 2016, 34(10): 1037–1045. [DOI]

(责任编辑: 赵要凤)