使用多个模型计算蛋白质结合亲和力的共识算法

Ayşenaz Ezgi Ergin 和 Deniz Turgay Altılar

计算机工程系，伊斯坦布尔技术大学，伊斯坦布尔，土耳其

摘要

主要组织相容性复合体（MHC）分子能够结合肽并将其呈递于细胞表面，在细胞介导的免疫中发挥着重要作用。随着多年来数据库和技术的发展，肽结合亲和力计算研究取得了显著进展。目前已开发出多种技术来预测肽与 MHC I 类的结合。由于 MHC I 类本身的特性，大多数相关研究都能取得更好的性能，且还有更多进展。考虑到不同方法和技术的使用，以及相似方法在不同蛋白质上的应用，研究人员根据蛋白质肽的结合亲和力进行了分类。为此，研究采用 MHCFlurry、NetMHCpan、NetMHC、NetMHCCons 和 ssm-pmbec 对 MHC I 类进行了研究。在本论文范围内开展的这些模拟中，各模型之间没有表现出整体优势，但已发现它们各自在不同方面优于其他模型。根据模型的组合使用方式，最佳结果可能有所不同。关键在于了解数据并采用合适的模型。但即便如此，差异也不会特别显著。由于共识方法与模型直接相关，模型越好，结果也越好。Xix

关键词

MHC I，HLA I，结合亲和力，mhcflurry，netmhcpan，netömhc，netmhccons，ssmpmbec

1. 引言

在群体层面，MHC I类基因具有显著的等位基因多样性；数百种已知的HLA等位基因各自与严格且可能独特的肽结合——

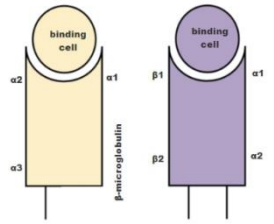


Figure 1.1 : Basic Structures of MHC Class I and MHC Class II

结合偏好。尽管肽能否被呈递的首要选择性要求是与 MHC I 类具有高亲和力结合，但抗原呈递通路中的其他步骤预计也会产生显著的次要影响。预测 MHC I类呈递肽是疫苗开发以及传染病、自身免疫和癌症研究中的一项重要技术。[1]当 T细胞识别并结合肽-MHC复合物时，可能会引发免疫反应，并使受累细胞裂解。因此，抗原肽与 MHC分子的结合是细胞免疫的关键阶段，而了解这一过程背后的模式具有

DOI：10.5121/ijab.2023.10101

改善人类健康的巨大潜力。[2] 多年来，人们采用了各种方法来计算蛋白质结合亲和力，而在此过程中遇到的困难促使人们

不同方法的推导。起初，人们更倾向于采用实验方法，但这些方法所需的实验设备价格昂贵，还耗费大量人力和时间，远非理想，因此促使人们转而利用计算能力和技术。最初，在这一领域的方法不断发展时，生物信息学面临的最大障碍是：小型数据集难以用于预测，或者无法确信预测结果是否正确。众所周知，数据集在计算蛋白质结合亲和力时占据非常重要的地位，其可靠性与数据集直接成正比。

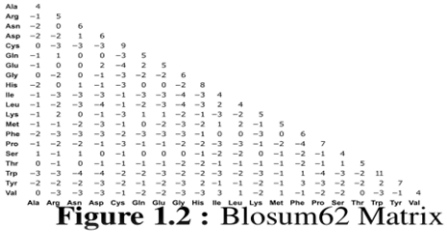


Figure 1.2 : Blosum62 Matrix

尽管如今我们将随后出现的应用归类为基于评分的方法，但在这些应用最初提出时，根据所采用的方法，可以将其分为三类：经验评分函数、基于知识的方法以及定量构效关系。[3] 在之后使用的方法中，人们尝试引入机器学习方法。由于缺乏数据集，早期研究人员认为需要使用更多特征，而过拟合正是他们面临的最大问题。这些技术的精确度取决于现有数据量，而这些数据定义了MHC分子的结合特异性。

2. 背景

2.1 MHC分子及其一般特性

HLA是人类中存在的MHC基因复合体形式。然而，MHC是一种基因复合体，可在脊椎动物所有有核细胞中表达表面抗原。MHC分子的主要功能是将抗原呈递给T细胞，使T细胞能够识别外源性抗原，并将其与自身抗原区分开来。MHC分子将细胞内当前蛋白质的相关信息传递至细胞表面，使免疫系统能够在必要时作出反应，例如刺激细胞毒性T淋巴细胞摧毁病毒感染细胞，或通过辅助性T淋巴细胞激活B细胞。此外，MHC主要分为I类、II类和III类。然而，HLA与MHC的主要区别在于它们在脊椎动物中的形成方式。肽与HLA分子结合后形成肽抗原，随后被T细胞接受并识别，这构成了人体免疫系统的基础。肽结合核心的氨基酸序列决定了肽-MHC结合亲和力。[4]

白血病细胞中特异性表达或过表达的蛋白质所含的 MHC I 类限制性 T细胞表位，是可用于免疫策略的候选疫苗。I类主要组织相容性复合体分子具有两端封闭的肽结合位点，并优先结合长度为 8 至 9 个氨基酸、能够容纳在这一封闭结合区域内的肽。[6]

MHC II类分子倾向于结合长度为12至20个氨基酸的肽，这类肽可视为长肽。目标是准确预测肽-HLA结合，以便获得免疫系统所期望的应答。主要组织相容性复合体II类（MHC-II）分子位于抗原呈递细胞表面，可将源自细胞外蛋白质的肽呈递给辅助性T细胞。肽与MHC结合形成的复合物通常位于抗原细胞表面。此外，只有能被T细胞受体识别的肽才能触发免疫系统。基于此，识别T细胞表位可谓此项工作的重点之一。MHC-II分子是一种由α链和β链组成的异二聚体糖蛋白。由MHC α链和β链的残基形成的结合槽可容纳MHC-II分子呈递的肽。

肽结合沟两端开放，因此可以结合不同长度的肽。尽管MHC II类分子能够容纳不同长度的肽，但自然界中最常见的肽长度为13至25个残基。肽配体中主要与MHC结合沟相互作用的部分称为肽结合核心，通常由九个氨基酸组成。

由于其自身特性，MHC I 类的结合亲和力更容易预测。研究显示，其预测性能相当高，而且通常研究的表位也更多。图1.1：MHC I 类和 MHC II 类分子的基本结构。MHC I 类呈递的肽主要源自细胞质，而 MHC II类分子呈递的肽主要源自外源性抗原。肽与 HLA 分子结合是肽呈递和 T细胞识别的前提，而有用的 T细胞表位仅包含所有 HLA结合肽中的一部分。许多因素使 MHC II 类的结合亲和力难以预测。首先可以列出分子的多态性结构、可变肽长、PFR效应以及肽的准确定义。考虑到这些原因及其引发的各种次要因素，MHC I 类结合亲和力的预测变得更加困难。相比之下，与 MHC I 类结合亲和力预测相比，其预测性能明显较低。

2.2 蛋白质结合亲和力计算方法

Vari已使用多种模型，体现了计算技术的发展以及不同特征 s.
Thes根据其工作原理，将这些方法分为三类进行考察。

2.2.1 基于评分的方法

它们之间最大的差异在于用于计算结合评分的方法。这些数学性更强的方法并不适用于所有HLA等位型。图1.2：BLOSUM62矩阵。这些方法假设结构均一，以评分矩阵算法为核心，通过对其进行各种计算得出结合评分，再根据所设定的阈值方法／数值，判断其是否具有足够的结合亲和力。当前最受关注的方法包括SYFPEITHI、RankPep、PickPocket和MixMHCpred。此外，SYFPEITHI于1999年推出，可称为首个获认可的蛋白质结合亲和力计算工具。所有这些方法都使用位置特异性评分矩阵，其中一些方法在计算过程中也会用到BLOSUM矩阵。[4] 还应指出，基于评分的方法的局限性在于，其预测评分的计算方法相对简单，因为它们只处理序列相似性和模式等线性属性。

RankPep 与特定 MHC分子结合的肽具有相似的序列特征。因此，长期以来，人们一直利用序列模式预测与 MHC分子结合的肽。然而，事实证明，这类序列模式过于简单，因为仅凭模式中为数不多的残基，无法准确表示复杂的结合基序。[7] 为克服这一局限，RANKPEP 使用位置特异性评分矩阵（PSSM）或由一组已知可与特定 MHC分子结合的比对肽构建的谱，来预测 MHC-肽结合。要使 PSSM 能够有效预测 MHC-肽结合并准确描述结合基序，肽必须按照结构和/或序列特征进行比对 [8]。具体而言，RankPep 使用谱系数对所有长度与 PSSM 相同的潜在片段进行评分。然而，仅凭这些评分不足以判定某条肽是否可能是结合肽。为更准确地识别可能的结合肽，需要对生成该谱的比对中包含的所有肽序列进行评分，并将所得分数设为结合阈值。[9]MHC I 和 MHC II 分子以相似但不同的方式结合肽，因此，MHC I 和

MHC II 配体的比对结果需符合肽与相应 MHC 类别的结合模式。RankPep 可通过“http://imed.med.ucm.es/Tools/rankpep.html”在线使用，并向所有人免费开放。

SYFPEITHI包含MHC I 类和MHC II 类，且向公众开放。任何人都可以通过在线网址 http://www.syfpeithi.de/ 使用该工具；其数据库仅包含已发表的数据。SYFPEITHI 的设计旨在不断纳入有关 MHC 肽基序、MHC 配体、T细胞表位以及 MHC分子氨基酸序列的数据。可以说，在 BIMAS 之后，这是生物信息学领域计算蛋白质结合亲和力所迈出的最初几步之一。该方法构建一个二维数据数组，其中字母代表行索引，口袋编号代表列索引。从第一行索引开始，将序列拆分，并计算每个分子所含行索引分数的总和。随后重复该方法，直至处理完整个序列。最终得出的整数值可用于判断结合是否适合、是否不适合，或是否具有中等结合亲和力。[10] 根据至少80%的可信度标准，SYFPEITHI 的结合亲和力预测会列出最有可能由 MHC分子呈递的肽，以获得最符合要求的表位；据此可以认为，自然呈递的表位可能是所有预测肽中得分最高的前2%。

2.2.2. 机器学习方法

随着机器学习方法的普遍发展以及数据量的增加，这些方法已被用于蛋白质-肽结合亲和力研究，使非线性模式也能应用于计算。此外，利用机器学习技术的科学计算库在应用方面的最新进展，为部署用于预测蛋白质结合亲和力的集成计算工具铺平了道路。蛋白质-配体复合物的原子坐标可用于预测结合亲和力。得益于这些新的计算工具，研究人员得以将广泛的机器学习方法应用于蛋白质-配体相互作用研究。这些机器学习方法中最重要的环节，是利用监督式机器学习技术、卷积神经网络和随机森林等技术训练新的计算模型。[12] 基于机器学习的方法利用训练模型，根据提取的代表性特征生成评分，从而判定某种肽是结合肽还是非结合肽。尽管采用的方法和算法各不相同，基于机器学习的蛋白质结合亲和力计算工具基本上都包含四个步骤。[13]

- 第一步是获取一个训练数据集，其中包含经过实验验证、已确认能够结合的等位基因和肽。在此步骤中进行预测之前，找到经过验证的同种异型至关重要。

- 建立适用于所用肽和 HLA 等位型特性的特征。需要注意的是，肽和 HLA 的表现可能大不相同。还应注意，肽具有各自独特的性质，HLA分子也有其独特的性质。此外，两者的共同作用也可能有所不同。在估算结合亲和力时，应审慎而准确地进行特征选择。通常，这一步需要一定的生物学知识；多年来，许多研究人员都发现，围绕这一主题开展多学科研究是一项重大挑战。例如，对于同一条肽，MHC I类分子会表现出强结合，而 MHC II 类分子则可能形成非常弱的结合。

- 完成训练数据集和必要的特征编码后，应选择最合适的机器学习算法，并使用该算法创建模型，再对模型进行

训练。这些算法可以是传统的机器学习算法，例如 RELIEF算法、支持向量机和线性回归；也可以是更先进的机器学习方法，例如卷积神经网络（CNN）、深度置信网络（DBN）和自编码神经网络。CNN 是最容易应用于计算机科学以外学科并进行泛化的深度学习算法之一，也是最常用的方法。CNN 是一种采用前馈结构的神经网络，与传统的人工神经网络相似。[14]

• 应对模型进行优化并评估其性能。这里最大的问题之一是过拟合。如果模型开始过度记忆用于训练的数据集，或者训练集过于单一，那么过拟合的风险就很高。如果将测试数据交给这个模型，尽管它在训练集上得分很高，我们很可能会得到非常低的分数。这是因为模型记住了训练集中的情况，并在测试数据集中寻找这些情况。由于哪怕稍有变化，都无法找到任何记忆过的案例，因此测试数据集上的预测评分可能会非常差。我们可以使用各种算法来避免过拟合。计算蛋白质结合亲和力估计时，最常用的算法是反向传播。[15]

本研究探讨了基于神经网络的工具。一般而言，HLA-肽结合预测模型中的神经网络采用分层前馈结构。典型的多层前馈神经网络由输入层、隐藏层和输出层组成。每一层都可以包含代表信号的神经元或单元。

NETMHCII 的首要特点是它属于等位基因特异性方法。它只能预测训练数据集中包含的 MHC 分子的结合亲和力，并且需要针对每种 MHC 分子分别进行训练。由于这种方法具有等位基因特异性，因此在可用数据集规模较大时应优先采用。[16] 肽和 MHC 伪序列使用 BLOSUM50 矩阵进行编码，而侧翼区则使用结合核心两端最多三个氨基酸窗口内的平均 BLOSUM评分进行编码。NetMHCII 使用 50 个网络，选取其中表现最好的 10 个，再对它们应用共识算法。[17]

NetMHCIIPan 是一种泛特异性技术，只要已知蛋白质序列，就能预测任意 MHC 分子。NetMHCIIPan 通过结合 MHC-II分子的信息来实现泛特异性，这些信息由被认为对肽结合至关重要的残基构成的伪序列表示。NetMHCIIPan 可构建一个通用网络。泛特异性方法的性能与最近邻有关。[16] [18]

2.3. 共识方法

共识技术的简洁性和稳健性至关重要。组合多种技术时，如果它们在所评估的条件下单独使用时表现明显优于各个方法，就会将其纳入最终方法的定义中。[19]此外，已有证据表明，由两种或更多独立方法组合而成的共识技术能够提高预测准确率。方法如此之多，非专业用户可能很难为预测与特定 MHC分子的结合选择最佳方法。顾名思义，这些方法是将多种方法组合起来，并通过共识算法作出决策。工具种类繁多且都得到应用，主要是因为它们在不同数据及其分布下表现各异。这里重要的共识模型旨在利用多种工具，获得更可靠的结果。这实际上是共识算法的一般原则，即力求遵循多数决策。共识算法不仅会采纳多数票的意见，也会接纳对所有参与者都有利的方案。因此，这始终是网络的双赢之举

。这一思维原则在并行计算中广泛应用，也让人推测，共识方法或许可以通过并行化来提升性能。尽管这与并行化十分相近，但仍需注意前馈架构。我认为可以用分布式架构取代分层架构，从而提升性能。例如，NetMHCCons 会使用 3 种不同的方法（2 种基于神经网络，1 种基于位置特异性评分矩阵），以适用于相应的 HLA 等位型。该应用也利用了最近邻检测。[13] 另一项研究发现，将 NetMHCII 和 NetMHCpan 这两种内部特性不同的并行方法结合使用，所得表现优于单独使用其中任何一种方法。[16]

NetMHC 是一种基于人工神经网络（ANN）的等位基因特异性方法，利用 94 种 MHC I 类等位基因进行训练。3.4 版作为 NetMHCcons 1.1 的组成部分使用。NetMHCpan 是一种泛特异性人工神经网络方法，使用超过 115,000 条定量结合数据进行训练，涵盖 120 多种不同的 MHC分子。2.8 版作为 NetMHCcons 1.1 的组成部分使用。PickPocket 方法基于矩阵，依赖于 MHC分子之间受体结合口袋的同源属性。该方法使用 94 种不同的 MHC等位基因进行训练。在 PickPocket 1.1 版中，口袋库矩阵由 SMMPMBEC 方法生成。NetMHCcons 1.1 服务器可为长度为 8 至 15 个氨基酸的肽生成预测结果，也允许用户选择多种长度。服务器支持两种提交类型：肽列表或 FASTA 格式的蛋白质序列。用户可以从较长的等位基因列表中选择所需的 MHC分子，也可以上传感兴趣的 MHC 蛋白质序列。用户可以选择基于预测亲和力（IC50）或百分位排名设定阈值，以定义强结合肽和弱结合肽。输出中会标示强结合肽和弱结合肽。输出还可按预测结合亲和力排序，并根据用户指定的阈值进行筛选。[19]

3. 工作流程

尽管用于计算蛋白质结合亲和力的模型采用不同的原理，但它们通常必须符合特定背景，因此不可能得出彼此差异很大的数值。此外，在标注结合强度（强、中、弱）时，预计这些模型的联合判断不会经常出现截然不同的结果。我们使用四种不同的模型进行了数值估计和标注研究。

模型 01-09；在本研究中，四分位距取数据的 0.1 至 0.9 分位数范围。忽略其余估计值，并计算该聚类的均值。模型 02-08；在本研究中，四分位距取数据的 0.2 至 0.8 分位数范围。忽略其余估计值，并计算该聚类的均值。

模型 035-065；在本研究中，四分位距取数据的 0.35 至 0.65 分位数范围。忽略其余估计值，并计算该聚类的均值。该模型最大的问题是，当数据不呈正态分布时，无法进行任何计算。

模型中位数；在此模型中，完成所有计算后，取预测值集合的中位数。

3.1 使用的模型

理论和实践中获得的所有知识都表明，不存在单一的优秀模型。模型之间的优缺点会直接影响结果。研究表明，共识模型的表现优于其他模型。其主要原因是，单独使用时看似不足的模型，在相互结合后开始产生更好的结果。我们的目标是让这些模型协同工作，从而获得最准确的结果。协同工作时，会根据输入序列为模型分配相应的系数，或者完全不使用该模型。例如，NetMHC训练数据集与序列之间的关系会直接影响结果。通过查看训练数据集，可以判断是否能够使用该模型，或者在使用时应让它对结果产生多大影响。

3.1.1 MHCFlurry

该程序采用全新的架构和肽编码技术，构建等位基因特异性神经网络。在以亲和力测量数据进行训练时，MHCFlurry在一项针对质谱发现的配体的基准测试中表现优于传统预测工具 NetMHC 4.0 和 NetMHCpan 3.0，尤其是在非九肽方面。在一项规模有限的亲和力测量基准测试中，该已发表的预测工具展现出与成熟工具相当的准确度，其中包括近期发布的 NetMHCpan 4.0；模型选择使用了质谱数据集。MHCFlurry 的预测速度超过每秒 7,000 次，比 NetMHCpan 4.0 快 396 倍。[21] MHCFlurry 会分别预测依赖 MHC 等位基因和不依赖 MHC 等位基因的效应。针对现有的 MHC I 类配体数据（包括亲和力测量数据和质谱数据集），该算法首先训练了一个新的泛等位基因 MHC I 类模型，作为针对依赖 MHC 等位基因效应的独立预测工具。为减少依赖 MHC 等位基因效应预测工具学习单个抗原加工预测信号的倾向，设计者采取了多项措施，其中之一是在训练数据中使用体外亲和力测量数据；这类数据在很大程度上独立于单个抗原加工预测。[22]

MHCFlurry 是一款开源工具，使用起来非常简单。所有人都可以通过 https://github.com/openvax/mhcflurry 访问它；它完全以 Python 编写，源代码也向公众开放。MHCFlurry 实现了 I类肽-MHC结合亲和力预测。最新版本包含泛MHC I预测器，可用于任何已知的 MHC等位基因。MHCFlurry 使用 TensorFlow神经网络库，并在 Python 上运行。 [22] 单个生物分子与其配体/结合伙伴之间结合关系的强度称为结合亲和力。平衡解离常数（KD）常用于测量和报告结合亲和力，它可量化并对生物分子相互作用的强度进行排序。KD 值越低，配体与其靶标的结合亲和力越高；KD 值越高，靶分子与配体之间的吸引和结合越弱。 [23]MHCFlurry 会输出估算的亲和力值，并以预测值的形式返回以 nM 为单位的亲和力（KD）。模型集成中第 5 至第 95 百分位的预测范围由预期低值和预测高值字段给出。[21] 由此定义可知，亲和力值与结合强度成反比。在此阶段，使用这款可通过命令行轻松操作的工具，计算了 3,465 条长度为 8 的肽在 HLA-A0201 和 HLA-A0301 等位基因上的 mhcflurry-binding-affinity 和 mhc-binding-affinity-percentage。对所得结果进行聚类时，将 mhcflurry-affinity-percentile 评定为：70–100 为弱，45–70 为中等，0–45 为强结合。下表列出了其中一部分数据。针对该等位基因测试了大量随机选取的肽，mhcflurry 亲和力百分位数表示该肽的亲和力预测值在这些肽中的百分位排名（范围为 0–100）。数值越低，结合越强。通常采用 2% 的阈值。[21]

MHCFlurry 是一项开源资源。可通过 https://github.com/openvax/mhcflurry 获取源代码和实现。本研究使用默认特征。在未更改任何参数的情况下，将各种肽和等位基因作为输入进行测试。

3.1.2 NetMHC

该模型最显著的特点是由50个神经网络组成。模型从50个神经网络中选取结果最佳的前10个模型，并依据皮尔逊相关系数筛选这10个模型。蛋白质结合亲和力所采用的相似性度量会直接影响结果。模型使用BLOSUM62相似性矩阵（其他模型也使用该矩阵），并通过稀疏编码得出正确结果。[17]NetMHC在单个等位基因层面学习到的长度偏好，大致反映了用于训练神经网络的数据分布。根据一项先前研究，对于HLA-A*02:01和HLAB*07:02，约30%的九肽和十肽在测量中表现为结合肽；然而，能够与这些分子产生显著测量亲和力的八肽和十一肽相对较少。[24] 简要概括该模型的特点：它使用人工神经网络进行预测。训练数据集中包含81种不同的人类等位基因和41种不同的动物等位基因。模型可以预测任意长度的肽，但建议使用9个氨基酸，且不超过11个。由于该模型针对特定等位基因，因此遵循这一建议对于确保准确性至关重要。[25]

NetMHC 可从 https://services.healthtech.dtu.dk/service.php?NetMHC-4.0. 提供学术用途下载。 .2.2.1 实现 它为非学术用户提供了一个电子邮件地址，以便与其联系。它还通过在线服务器为希望使用该工具的用户提供直接输出。此外，数据集也可以用同样的方式下载。本文使用 NetMHC Linux 4.0 版开展研究，并基于 IEDB 数据集取得进展。数据集可从 http://tools.iedb.org/mhci/download/ 获取。

3.1.3 NetMHCpan

NetMHCpan-4.1 服务使用人工神经网络，预测肽与任意已知序列的 MHC分子之间的结合。该方法使用了 850,000 多条可定量的结合亲和力数据和质谱洗脱配体肽进行训练。BA数据涵盖来自人类和不同动物的 170 种 MHC分子。EL数据集包含来自人类的 177 种 MHC分子。此外，用户上传完整长度的 MHC 蛋白质序列后，即可获得针对任意自定义 MHC I类分子的预测。该服务可预测任意长度的肽。[2] 由于泛特异性方法取决于算法识别肽与 MHC序列之间关系相似性的能力，因此显而易见，当查询的 MHC分子由与结合特异性特征密切相关的 MHC分子表示时，该方法的效果会更好。 [26]泛特异性方法依赖神经网络理解肽与 HLA伪序列相互作用的广泛特征，并以结合亲和力为依据。只有训练集数据中具有多态性的相互作用，才能促进神经网络的学习过程。由于训练数据在扩展MHC位点缺乏多态性，NetMHCpan 方法无法从此类扩展伪序列映射中提取信息。

NetMHCPan 提供学术用途下载：https://services.healthtech.dtu.dk/service.php?NetMHCpan-4.1。非学术用户可通过其提供的电子邮件地址进行联系。它还提供在线服务器，供用户直接获取输出。此外，数据集也可以通过相同方式下载。本研究基于 IEDB数据集开展。数据集可在 http://tools.iedb.org/mhci/download/ 获取。

3.1.4 NetMHCCons

总体来看，各模型相较于彼此都各有优缺点。因此，旨在通过使用共识模型来突出优势并尽量减少劣势。共识模型通过多个模型获得结果。在该模型中，采用了三种不同的方法，分别是 NetMHCpan、NetMHC 和 PickPocket。PickPocket 是一种基于序列的模型，也就是说，它基于线性特征进行分析。

当某个等位基因存在于训练数据集中时，所有三种方法及其组合均可用于确定共识方法。在确定最佳共识技术时，某些方法比其他方法更为重要，因为它们通常能够达到最高性能。

NetMHC提供学术用途下载：https://services.healthtech.dtu.dk/service.php?NetMHCcons-1.1。非学术用户可通过其提供的电子邮件地址进行联系。它还为希望使用该工具的用户提供在线服务器，可直接输出结果。此外，也可以通过相同方式下载数据集。本研究使用了IEDB数据集。数据集可从 http://tools.iedb.org/mhci/download/ 获取。由于该模型是共识模型，因此还需要加载其使用的 PickPocket、NetMHC和NetMHCCons模型。

3.1.5 SSMPMBEC

针对肽构建了一种名为 PMBEC 的氨基酸相似性矩阵。该矩阵的一个显著特点是无法实现相反电荷氨基酸之间的替换。[28] 这很可能是肽的一般特性。基于这一思路，研究采用 PMBEC 矩阵来构建 SSM 模型。长期以来，使用 NetMHC 开展的研究中，各项范围并无明显差异。

SSMPMBEC 实现也可在 https://github.com/ykimbiology/smmpmbec 获取。

4. 结果

结果清楚地表明，无法选出一种适用于所有情况的通用模型。选择模型时，所用数据集是最重要的考量因素之一。结合使用时，亲和力计算的结果更加令人满意。本研究使用的数据集为 IEDB训练数据集。通过多幅图展示了预测与模型的拟合情况。正如预期，11肽的结果差异最大。与其他肽相比，9肽和10肽的相容性更好。我们还注意到，每种模型都有一定的偏差范围。通过剔除所产生的离群值，所用模型在真实模型附近呈现出增长趋势。这表明，这些方法是围绕真实结果构建的。图表中的数据经过归一化处理。由于所构建模型的结论与所用模型一致，所用模块的准确性会直接影响构建出的模型。总体而言，错误标注问题出现在极限值处。

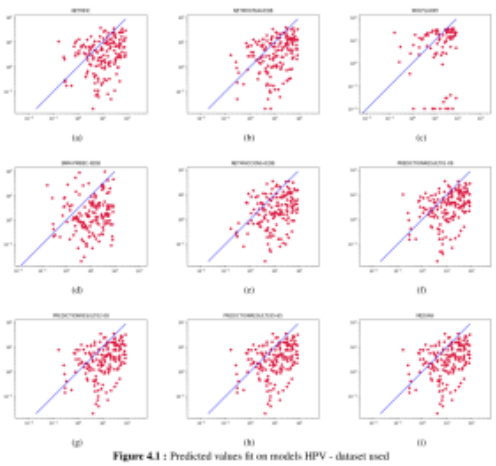


Figure 4.1 : Predicted values fit on models HPV - dataset used

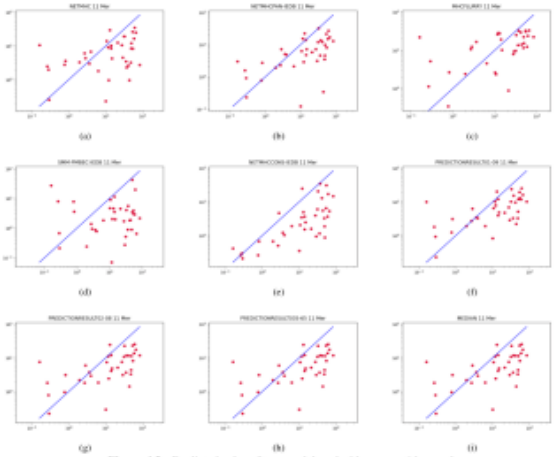


Figure 4.2 : Predicted values fit on models only 11-mer peptides used

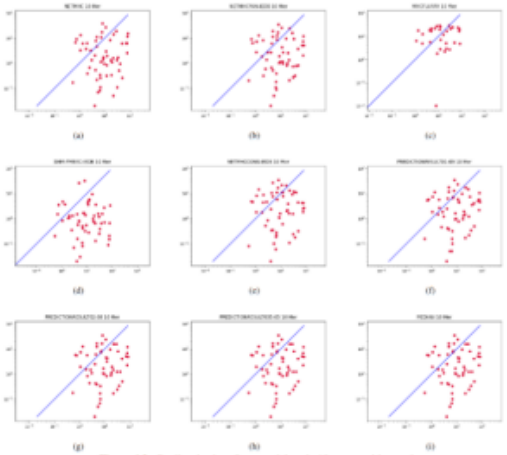


Figure 4.3 : Predicted values fit on models only 10-mer peptides used

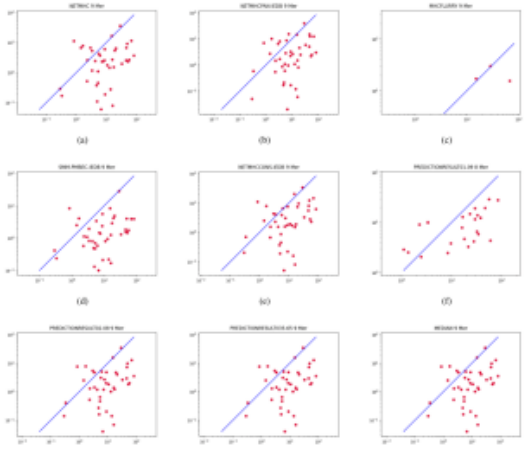


Figure 4.4 : Predicted values fit on models only 9-mer peptides used

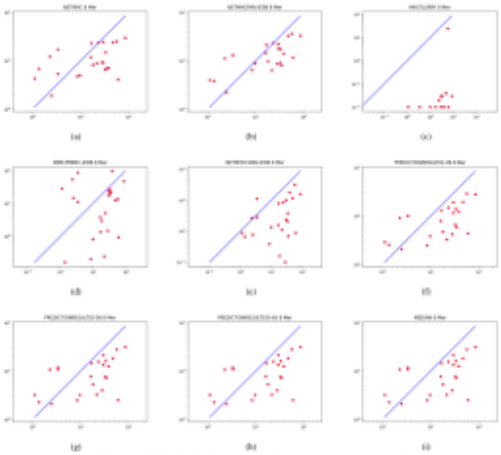


Figure 4.5 : Predicted values fit on models only 8-mer peptides used

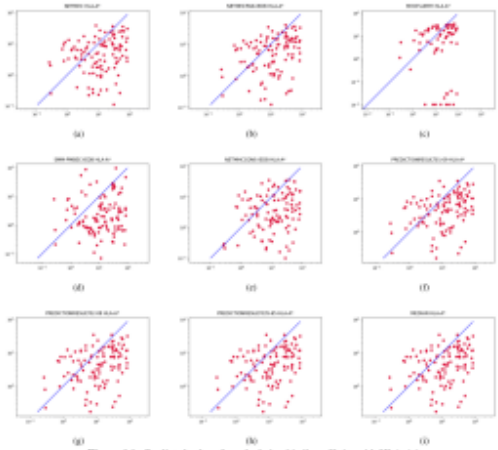


Figure 4.6 : Predicted values for calculation binding affinity with HLA-A*

5. 结论

我们无法将所用模型归类为高效或低效。尽管共识模型通常能取得更好的结果，但给定的数据和训练集会直接影响结果。本研究面临的最大障碍是数据有限，且当前资源有限。确实存在某些数据条目，使得各模型分别在其中表现最佳；即使进行比较，各模型之间也没有太大差异。由于 MHCFlurry 和 SMMPMBEC 的结果偏差较大，在已知这两个模型的数据集时，可以使用其他模型；而当数据集的不确定性较高时，使用其他模型则可能得到更可靠的结果。

数据与模型改进直接相关这一事实，实际上使我们意识到，改进模型时所采用的数据整理方式更为重要。基于评分矩阵得出这些推论。因此，旨在通过深度学习创建评分矩阵，以改进模型。

结果清楚地表明，无法选定一个通用模型。用于选择模型的数据集是最重要的考量因素之一。结合使用时，亲和力计算结果更令人满意。此处使用的数据集是IEDB训练数据集。预测与模型的拟合情况通过多幅图表展示。正如预期，差异最大的是11肽。与其他肽相比，9肽和10肽的兼容性更好。我们还注意到，每个模型都存在偏差范围。移除生成的离群值后，所用模型的结果整体上更趋近于真实模型。这表明，这些方法是围绕真实结果构建的。图表中的数据经过归一化处理。我们尝试通过在不同聚类中检查数据，理解数据的含义，并了解模型与数据的相关性。首次聚类时，根据等位基因长度将数据集分为4部分。研究了8肽、9肽、10肽和11肽聚类中，模型值与实际值的重叠程度。我们考察了所用模型的表现，以及基于整个数据集构建的方法。图4.1是在解析整个数据集之前生成的。在这些模型中，最引人注目的或许是MHCFlurry：它要么收敛到最准确的值，要么直接得出与之偏差很大的值。在所考察的模型中，ssm-pmbec给出的结果偏差最大，而最接近真实结果的模型则是NetMHCCons 27。创建的共识模型，其性能与所用模型成正比。虽然预期模型中的离群值比例变化会带来改进，但最佳结果出现在01-09模型中。这是因为，在其他模型中被视为离群值的结果影响了实际模型的结果，使其更接近真实情况。换言之，01-09模型中那些在其他模型里被视为离群值的数据，有助于改善结果，也就是降低了误差范围。尽管MHCflurry的结果有时非常接近模型，有时又相距甚远，但在各个聚类中可以观察到，随着等位基因长度缩短，结果会离真实值越来越远。预期最佳结果应来自9肽或10肽。初步研究表明，这些模型通常在9肽和10肽上表现更好。与9肽和10肽相比，模型对10肽的预测更为成功，这直接影响了本研究提出的共识方法，使其对10肽的预测结果更加趋同。最显著的是MHCFlurry的结果。这些结果表明，单独使用该模型可能产生严重误导，因为其结果要么过于接近，要么偏离甚远。虽然由于所用数据中HLA-A*出现频率较高，人们可能认为以HLA-A*为输入的模型表现更好，但从偏差角度来看，两者之间并无显著差异

参考文献

[1] Peters, B., Nielsen, M. and Sette, A. (2020). T细胞表位预测，《免疫学年度评论》，38(1), 123–145, https://doi.org/ 10.1146/annurev-immunol-082119-124838. [2] Reynisson, B., Alvarez, B., Paul, S., Peters, B. and Nielsen, M. (2020). NetMHCpan-4.1和NetMHCIIpan-4.0：通过同时进行基序解卷积并整合质谱MHC洗脱配体数据，改进MHC抗原呈递预测，《核酸研究》，48(W1), W449–W454, https://doi.org/10.1093/nar/gkaa379. [3] Yugandhar, K. and Gromiha, M.M. (2014). 基于氨基酸序列的蛋白质-蛋白质结合亲和力预测，《生物信息学》，30(24), 3583–3589, https://doi.org/10.1093/bioinformatics/btu580. [4] Gomez-Nunez, M., Pinilla-Ibarz, J., Dao, T., May, R.J., Pao, M., Jaggi, J.S. and Scheinberg, D.A. (2006). 肽结合基序预测算法与实验结合相对应，293–1298. [5] Paul, S., Weiskopf, D., Angelo, M.A., Sidney, J., Peters, B. and Sette, A. (2013). HLA I类等位基因与不同大小、亲和力和免疫原性的肽结合库相关，《免疫学杂志》，191(12), 5831–5839, https://doi.org/10.4049/ jimmunol.1302101. [6] Biddison, W.E. and Martin, R. (2000). MHC I类和II类分子的肽结合基序，《免疫学现行实验方案》，36(1), https: // doi.org/10.1002/0471142735.ima01is36. [7] Lundegaard, C., Lund, O., Buus, S. and Nielsen, M. (2010). 主要组织相容性复合体I类结合预测作为表位发现工具，《免疫学》，130(3), 309–318, https://doi.org/10. 1111/j.1365-2567.2010.03300.x. [8] Karosiene, E., Lundegaard, C., Lund, O. and NetMHCcons, N.M. (2011). 一种用于主要组织相容性复合体I类预测的共识方法。《免疫遗传学》。2012 Mar;64(3):177-86, EpubOct, 20. [9] Reche, P.A., Glutting, J.P. and Reinherz, E.L. (2002). 使用谱基序预测MHC I类结合肽，《人类免疫学》，63(9), 701–709, https://doi.org/10.1016/s0198-8859(02) 00432-9.[10] Rammensee, H.G., Bachmann, J., Emmerich, N.P.N., Bachor, O.A. and Stevanović, S. (1999). SYFPEITHI：MHC配体和肽基序数据库，《免疫遗传学》，50(3-4), 213–219, https://doi. org/10.1007/s002510050595. [11] Schuler, M.M., Nastke, M.D. and Stevanovi, S., (2007). SYFPEITHI，《分子生物学方法》，Humana Press, pp.75–93. [12] Bitencourt-Ferreira, G. and de Azevedo, W.F., (2019). 用于预测结合亲和力的机器学习，《分子生物学方法》，Springer纽约, pp.251–273. [13] Mei, S., Li, F., Leier, A., Marquez-Lago, T.T., Giam, K., Croft, N.P., Akutsu, T., Smith, A.I., Li, J., Rossjohn, J., Purcell, A.W. and Song, J. (2020). HLA I类肽结合预测生物信息学工具的综合综述与性能评估，《生物信息学简报》，21(4), 1119–1135, https://doi.org/10.1093/bib/bbz051. [14] Zhao, J., Cao, Y. and Zhang, L. (2020). 探索蛋白质-配体结合位点预测的计算方法，《计算与结构生物技术杂志》，18, 417–426, https://doi.org/10.1016/ j.csbj.2020.02.008. [15] Gill, N. and Singh, S. (2011). 使用布尔代数和模糊逻辑进行多序列比对：比较研究，《国际计算机技术与应用杂志》，02. [16] Jensen, K.K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J.A. and Yan, Z. Nielsen, M. (2018). 改进预测肽与MHC II类分子结合亲和力的方法，《免疫学》，154(3), 394–406. [17] Andreatta, M., Karosiene, E., Rasmussen, M., Stryhn, A., Buus, S. and Nielsen, M. (2015). 通过改进结合核心识别，准确进行肽-MHC II类结合亲和力的泛特异性预测，《免疫遗传学》，67(11-12), 641–650, https://doi.org/10.1007/ s00251-015-0873-y. [18] Karosiene, E., Rasmussen, M., Blicher, T., Lund, O., Buus, S. and Nielsen, M. (2013). NetMHCIIpan-3.0：一种通用的泛特异性MHC II类预测方法，涵盖全部三种人类MHC II类同种型HLA-DR、HLA-DP和HLA-DQ，《免疫遗传学》，65(10), 711–724, https://doi.org/10.1007/s00251-013-0720-y. [19] Karosiene, E., Lundegaard, C., Lund, O. and Nielsen, M. (2011). NetMHCcons：一种用于主要组织相容性复合体I类预测的共识方法，《免疫遗传学》，64(3), 177–186, https: //doi.org/10.1007/s00251-011-0579-8. [20] Alvarez, B., Reynisson, B., Barra, C., Buus, S., Ternette, N., Connelley, T., Andreatta, M. and Nielsen, M. (2019). NNAlign_MA MHC肽组42解卷积，实现准确的MHC结合

基序表征与改进的T细胞表位预测，《分子与细胞蛋白质组学》，18(12)，2459–2477，https://doi.org/10.1074/mcp.tir119. 001658. [21] O'Donnell, T.J., Rubinsteyn, A., Bonsack, M., Riemer, A.B., Laserson, U. 和 Hammerbacher, J. (2018). MHCflurry：开源 I类主要组织相容性复合体结合亲和力预测，《细胞系统》，7(1)，129–132.e4，https: //doi.org/10.1016/j.cels.2018.05.014. [22] O'Donnell, T.J., Rubinsteyn, A. 和 Laserson, U. (2020). MHCflurry 2.0：通过纳入抗原加工改进 MHC I类呈递肽的泛等位基因预测，《细胞系统》，11(1)，42–48.e7，https://doi.org/10.1016/j.cels.2020.06.010. [23] Du, X., Li, Y., Xia, Y.L., Ai, S.M., Liang, J., Sang, P., Ji, X.L. 和 Liu, S.Q. (2016). 深入了解蛋白质-配体相互作用：机制、模型与方法，《国际分子科学杂志》，17(2)，144，https://doi.org/10.3390/ijms17020144. [24] Andreatta, M. 和 Nielsen, M. (2015). 使用人工神经网络进行带缺口序列比对：在MHC I类系统中的应用，《生物信息学》，32(4)，511–517，https://doi.org/10.1093/ bioinformatics/btv639.[25] Richters, M.M., Xia, H., Campbell, K.M., Gillanders, W.E., Griffith, O.L. 和 Griffith, M. (2019). 新抗原临床应用的生物信息学表征最佳实践，《基因组医学》，11(1)，https:// doi.org/10.1186/s13073-019-0666-2. [26] Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Røder, G., Peters, B., Sette, A., Lund, O. 和 Buus, S. (2007). NetMHCpan：一种定量预测肽与任何已知序列的 HLA-A 和 -B 位点蛋白质结合的方法，《公共科学图书馆·综合》，2(8)，e796，https://doi.org/10.1371/journal.pone. 0000796. [27] Hoof, I., Peters, B., Sidney, J., Pedersen, L.E., Sette, A., Lund, O., Buus, S. 和 Nielsen, M. (2008). NetMHCpan：一种超越人类的 MHC I类结合预测方法，免疫遗传学，61(1)，1–13，https:// doi.org/10.1007/s00251-008-0341-z. [28] Kim, Y., Sidney, J., Pinilla, C., Sette, A. 和 Peters, B. (2009). 推导用于肽-MHC结合的氨基酸相似性矩阵，并将其作为贝叶斯先验应用，《BMC生物信息学》，10(1)，https://doi.org/ 10.1186/1471-2105-10-394。