PIBB 生物化学与生物物理进展
*Progress in Biochemistry and Biophysics*
2024, 51(10): 2532~2544
www.pibb.ac.cn

# 抗原表位预测工具的研究与发展现状*

李梓豪[1)] 汪 源[2)] 毛甜甜[2)] 曹志伟[1)**] 裘天颐[3)**]
([1)] 复旦大学生命科学学院，上海 200433；[2)] 同济大学生命科学与技术学院，上海 200092；
[3)] 复旦大学附属中山医院实验研究中心，上海 200032）

**摘要** 适应性免疫在抗原识别和人体免疫过程中起到十分重要的作用。本文综述了抗原表位预测工具的研究进展及其在疫苗设计和免疫治疗策略中的应用，突出了其重要性。通过分析B细胞和T细胞抗原表位的识别机制，本文阐释了表位的种类及其在免疫反应中的作用。进一步详细讨论了B细胞和T细胞抗原表位的预测工具，特别是它们如何利用支持向量机、随机森林、深度学习等不同算法来解析表位信息，并介绍了当前该领域的最新发展现状。最后，本文对抗原表位预测技术的未来发展趋势进行了展望。

**关键词** 抗原表位预测，适应性免疫，B细胞，T细胞，机器学习，深度学习
**中图分类号** Q-1 **DOI**：10.16476/j.pibb.2024.0351

免疫系统是人体抵御外来物质入侵的重要防御机制。它由体内的多种细胞和分子共同参与，形成天然的免疫屏障。免疫系统主要分为固有免疫（又称非特异性免疫）和适应性免疫（又称特异性免疫）两大类[1]。固有免疫是机体在种系发育和进化过程中形成的自然防御功能，其反应迅速，但缺乏特异性，且反应强度相对较弱。相反，适应性免疫具有高度特异性，尽管首次免疫识别时间较长，但在再次接触同一抗原时，其反应迅速且强烈。适应性免疫通常在固有免疫应答后发挥效应，其目的是清除抗原，并形成记忆细胞以防同一抗原再次入侵。

适应性免疫主要分为细胞免疫和体液免疫两个部分[2]。细胞免疫通常由T细胞参与，可以直接杀死被内源或外源性抗原感染的细胞，或激活免疫系统中的其他细胞参与反应；体液免疫主要由B细胞发挥作用，通过分泌抗体直接或间接消灭抗原，并留下记忆细胞应对相同抗原的下次入侵。适应性免疫过程十分复杂，涉及多个层次、多种组织和大量细胞，各层次之间存在一定的联系和自主性（图1）。

分子特异性识别是适应性免疫的关键核心问题。当两个或多个分子在相互识别并结合后，就会形成稳定的复合物。在细胞免疫中，T细胞受体会与主要组织相容性复合体（major histocompatibility complex，MHC）-多肽复合物结合；在体液免疫中，抗原会与抗体结合。这些都是外源性物质与机体应答分子相互识别并结合的过程。分子识别的关键在于免疫分子对外源性物质的线性或空间特殊区域——表位（epitope）的识别[3]。T细胞、B细胞和可溶性抗体对表位的识别是获得性免疫应答的核心，这种分子识别依次激活细胞免疫和体液免疫系统，进行人体适应性免疫应答，并起到清除内源性或外源性抗原物质的目的。

* 国家自然科学基金（32370697）资助项目。
** 通讯联系人。
曹志伟 Tel：021-65980296，E-mail：zwcao@fudan.edu.cn
裘天颐 Tel：021-64041990，E-mail：tianyi_qiu@fudan.edu.cn
收稿日期：2024-07-30，接受日期：2024-09-03

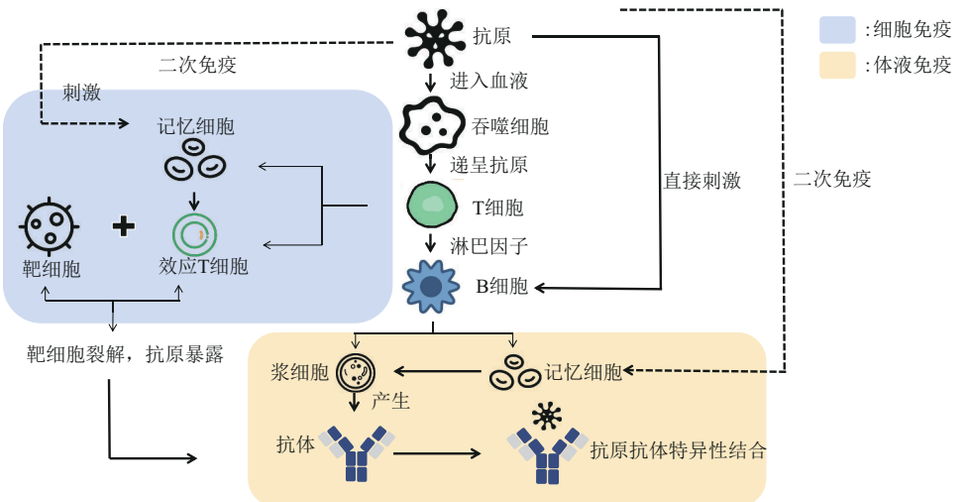



**Fig. 1　Schematic diagram of adaptive immunity**

**图1　适应性免疫示意图**

## 1　抗原表位

抗原表位是指抗原分子或细胞表面上的特定区域，通常是抗原表面上的5~8个氨基酸的残基，这些区域能够被免疫系统中的B细胞受体（抗体）或T细胞受体所识别和结合。这些表位通常是抗原分子上的特定结构，免疫系统能够通过识别它们来启动免疫响应。并且抗原表位在免疫系统的功能中扮演着至关重要的角色，因为它们能够触发对外来病原体或异常细胞的免疫反应，从而保护机体免受感染和疾病的侵害[4]。

传统的抗原表位发现方法主要依赖于实验技术，例如多肽扫描、X射线晶体学和核磁共振成像等[5]。这些技术虽然能够精确识别出蛋白质中的表位结构，但却面临着诸多挑战。例如，这些方法通常需要大量的时间和资源投入，尤其是在进行大规模抗原分析时。而且，传统方法往往依赖于高度纯化的蛋白质和特定的实验设置，这也限制了其在复杂生物样本中的应用。所以，开发高效的抗原表位预测工具显得尤为重要。这些计算工具能够利用生物信息学算法，快速预测可能的表位区域，大大减少实验工作的负担。通过结合生物信息学、免疫学和计算模型，研究者能够在更广泛的样本和条件下进行表位分析，从而加速疫苗设计、精准医疗和早期诊断的研究进程[6]。

### 1.1　B细胞抗原表位

B细胞抗原表位是指抗原分子上能够与B细胞受体或抗体特异性结合的区域，它们通常是暴露在抗原表面的氨基酸序列，是免疫系统识别和清除病原体的关键组成部分（图2）。这些表位可以分为线性表位和构象表位两类[7]。线性表位是抗原分子上的连续氨基酸序列，由相邻的氨基酸组成。它们在抗原的主链上连续排列，形成线性结构。这种表位在抗原分子的原始构象中即存在，与抗原的空间结构无关。B细胞受体和抗体的可变区域通过与线性表位的氨基酸序列形成互补的空间结构，从而实现特异性识别和结合。构象表位是抗原分子上的非连续氨基酸序列，通常由在原始构象中不相邻的氨基酸组成。它们的形成受到抗原的空间结构的影响，可能需要特定的蛋白质折叠或结合状态才能暴露。构象表位的识别往往需要抗原分子的结构改变或折叠状态的变化，因此对于某些抗原来说，它们可能仅在特定的条件下才能被识别和结合。随着目前对蛋白质的空间结构和折叠状态的研究不断加深，尤其是冷冻电镜技术的发展应用，越来越多重要的抗体表位被发现位于抗原分子空间上的特定结构域或折叠结构中，而非线性结构，因此构象表位研究的重要性日渐显现。

B细胞抗原表位作为体液免疫过程中的重要组成部分，具有以下几个重要特点[8]。a. 免疫原性。B细胞抗原表位能够引发机体免疫系统的免疫应答。当B细胞受体与抗原表位结合时，会激活B细胞并诱导其分化为浆细胞，产生大量的抗体。这种免疫原性使得B细胞抗原表位成为疫苗设计和免疫

治疗的重要靶点。b. 特异性。B细胞抗原表位通常具有高度特异性，能够与特定的B细胞受体或抗体发生相互作用。这种特异性是由B细胞受体或抗体的可变区域决定的，它们能够与抗原表位形成互补的空间结构，从而实现特异性的识别和结合。c. 表位多样性。由于B细胞受体和抗体的可变区域是由基因重排和体细胞突变产生的，因此B细胞抗原表位具有多样性，也就是一个抗原可能存在多个不同的表位，而一个B细胞受体或抗体也可能与多个不同的抗原表位结合。d. 表位密度。抗原分子上的表位密度可以影响免疫应答的强度和效率。一般来说，表位密度越高，免疫应答越强烈。因此，在疫苗设计中，可以考虑增加抗原上表位的密度，以提高疫苗的免疫原性。

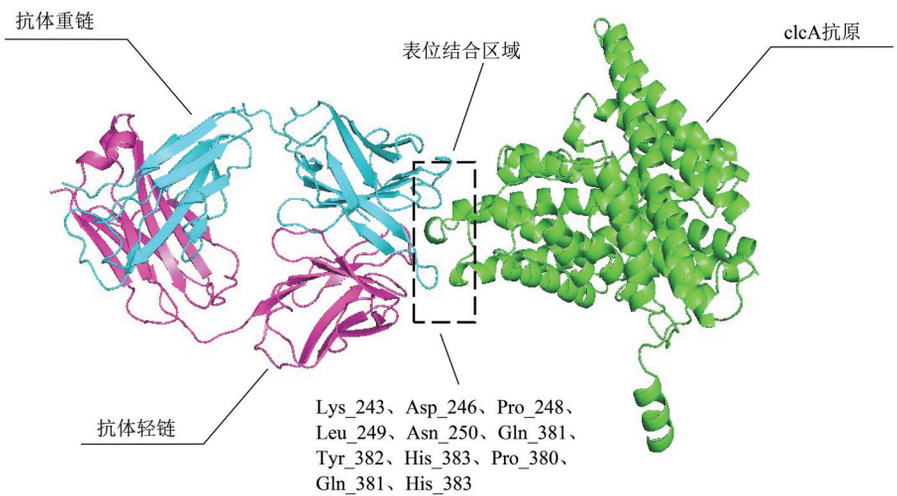



**Fig. 2　Schematic diagram of antigen-antibody interaction**

**图2　抗原抗体相互作用示意图**

图中结构为IgG蛋白抗体Fab片段与H(+)/Cl(−) exchange transporter clcA抗原（氯离子通道蛋白）结合相互作用的示意图（PDB ID：2HTK [9]），其中紫色为抗体的轻链，蓝色为抗体的重链，绿色为氯离子通道蛋白，黑色虚线框中标注为表位识别结合区域。

### 1.2　T细胞抗原表位

T细胞抗原表位与B细胞有所不同，它们通常由蛋白质抗原的特定线性片段组成，这些线性连续表位片段能够与T细胞受体（T cell receptor，TCR）的可变区域发生相互作用，并激活T细胞产生免疫应答反应。在细胞免疫过程中，T细胞抗原表位的产生涉及到多个关键步骤，包括抗原处理、抗原呈递、MHC分子结合以及T细胞受体的识别 [10]。

细胞免疫的第一步是抗原通过细胞内吞作用（例如吞噬作用或自噬作用）或蛋白酶体降解内源性抗原的过程来实现。在这个过程中，抗原被分解成较短的多肽片段，这些片段就是T细胞抗原表位。

处理后的抗原片段将会与MHC分子结合，并被MHC分子转运到细胞表面。MHC分子是T细胞抗原表位的主要载体，它们与抗原片段结合后形成MHC-抗原复合物。抗原片段与MHC分子的结合是高度特异的，通常取决于抗原片段的序列和MHC分子的型别。MHC分子包括类I和类II两种类型，它们分别呈递内源性和外源性抗原。在这个过程中，MHC分子会与抗原片段的特定结构区域相互作用，形成稳定的MHC-抗原复合物（pMHC）（图3）。

T细胞抗原表位的产生过程的关键步骤是TCR的识别。TCR是位于T细胞表面的受体蛋白，具有高度特异性。当T细胞受体与MHC-抗原复合物结合时，会激活T细胞并诱导其免疫应答。这种识别是通过TCR的可变区域与MHC-抗原复合物的特定结构相互作用实现的，形成T细胞抗原表位的识别，该步骤也是T细胞抗原表位预测的目标。

T细胞抗原表位是细胞免疫的基础，为T细胞识别和应对病原体、肿瘤细胞以及其他异常细胞提供了重要的机制，它具有以下几个显著特点 [11]。a. 免疫原性。T细胞抗原表位能够引发机体免疫系统的免疫应答。当TCR与抗原表位结合时，会激

活T细胞并诱导其分化为效应性T细胞，从而启动针对该抗原的免疫反应。b. MHC限制性。T细胞抗原表位的识别和结合通常受到MHC分子的限制。MHC分子能够将抗原表位呈递给T细胞，使得TCR能够与抗原表位及其MHC分子共同结合。因此，T细胞的抗原识别和结合通常依赖于特定的MHC类型。c. 可变性。T细胞抗原表位的可变性往往取决于抗原分子的性质以及免疫应答的环境条件。一些抗原可能具有多个不同的表位，而这些表位的识别和结合可能受到TCR的可变区域的影响。此外，在感染或肿瘤等异常状态下，抗原分子的表位可能会发生变异，导致T细胞的抗原识别产生改变。d. 表位限制性。每个T细胞抗原表位通常只能与特定类型的TCR结合。这种表位限制性意味着不同的T细胞群体可能会对不同的抗原表位产生反应，从而实现对多种抗原的识别和应对。

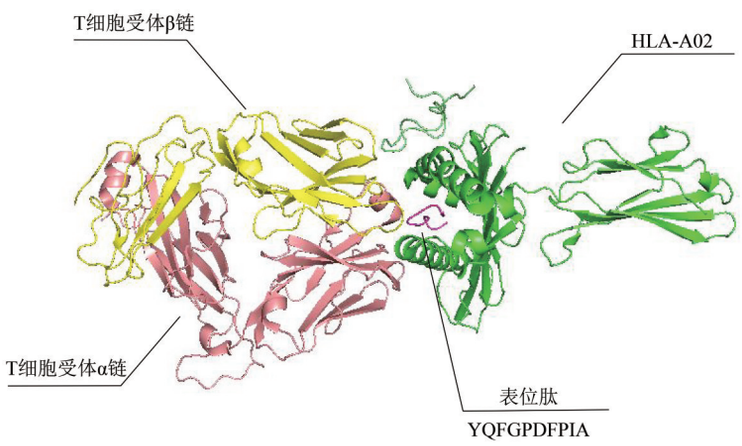



**Fig. 3　MHC−peptide−TCR interaction**

**图3　MHC−peptide−TCR 相互作用示意图**

图中结构为1E6 TCR与HLA-A02和表位肽结合相互作用的示意图（PDB ID：5C07[12]）。绿色为人I类组织相容性复合物HLA-A-2 α链；紫色为肽段，即T细胞表位；粉色为T细胞受体α链；黄色为T细胞受体β链。

## 2　B细胞抗原表位预测工具

B细胞抗原表位预测工具是一类用于预测蛋白质抗原分子中潜在B细胞表位的计算工具。这些工具的发展是为了帮助研究人员更好地理解免疫应答的机制、设计有效的疫苗和免疫治疗策略。

B细胞抗原表位预测的主要训练任务是一个二分类问题，即将蛋白质序列中的每个残基标记为可能的表位残基或非表位残基。在这个任务中，训练模型需要提取抗原蛋白的结构特征（一到三级结构、溶剂可及性、氨基酸侧链之间的距离等）和理化性质（氨基酸大小、电荷、疏水性等）作为特征描述符[13]，从而通过模型算法学习如何区分蛋白质序列中的每个残基构成B细胞表位的可能性。

根据训练数据以及输入数据类型的不同，B细胞抗原表位预测工具可以分为如下3类[14]：基于抗原序列的B细胞抗原表位预测工具、基于抗原结构的B细胞抗原表位预测工具和抗体特异性B细胞抗原表位预测工具。

### 2.1　基于序列的B细胞抗原表位预测工具

基于序列的B细胞抗原表位预测工具是利用蛋白质序列信息进行预测的一类工具。该类工具的预测仅依赖于抗原的一级序列，而不需要三维结构。训练数据通常包含已知的B细胞表位序列和非表位序列，通过提取这些序列的生化性质、二级结构信息、溶剂可及性等特征，再结合氨基酸在序列中的相对位置、上下游序列的信息等上下文信息，以描述氨基酸在序列中的环境，然后采用机器学习或深度学习算法，学习已知的B细胞表位和非表位序列之间的特征差异，从而预测新的蛋白质序列中可能的B细胞抗原表位[13]。表1展示了不同方法开发的基于序列的B细胞抗原表位预测工具，总结了目前基于序列的预测方法。

**Table 1 List of sequence-based B-cell epitope prediction tools**

**表1 基于序列的B细胞抗原表位预测工具列表**

| 工具名称 | 发表年份 | 训练集/测试集 | 特征 | 算法 | ROC-AUC（文献中） |
|---|---|---|---|---|---|
| Epitopia [15] | 2009 | 260/NA | 能够使用蛋白质的三维结构或线性序列预测免疫原性区域 | 朴素贝叶斯分类器 | 0.59 |
| CBTOPE [16] | 2010 | 187/52 | 使用了三种特征提取方法：二进制模式的轮廓（BPP）、物理化学模式的轮廓（PPP）和模式的组成轮廓（CPP） | 支持向量机（SVM） | 0.90 |
| BepiPred-2.0 [17] | 2017 | 155/5 | 使用了蛋白质体积、疏水性、极性、相对表面可及性（RSA）和二级结构（SS）等特征 | 随机森林 | 0.62 |
| BepiPred-3.0 [18] | 2022 | 1 466/NA | 使用蛋白质语言模型以提高预测的准确性。能够处理线性和构象B细胞表位的预测 | ESM-2蛋白语言模型 | 0.77 |
| epitope-1D [19] | 2023 | 154 899个数据点/30 980个数据点 | 基于图的蛋白质序列签名和有机体本体识别信息。提供可解释的机器学习分类器，可以解释预测决策 | 解释性增强机器（EBM） | 0.935 |

## 2.2 基于结构的B细胞抗原表位预测工具

基于结构的B细胞抗原表位预测工具利用蛋白质的三维结构信息进行预测。这类工具通常需要蛋白质的PDB（protein data bank）[20]数据作为输入，然后通过分析蛋白质的结构特征和分子相互作用信息，预测蛋白质中可能存在的B细胞抗原表位。

依赖于PDB中的抗原蛋白结构数据，目前已经发表了很多基于结构的B细胞表位预测工具。与蛋白质序列数据不同，PDB数据通常包含蛋白质的三维结构信息，包括原子坐标、残基之间的连接关系等，这些基于实验技术得到的复合物结构，可以用于分析抗体与抗原之间的相互作用。所以PDB数据具有较高的数据质量和多样性，这些数据能够为模型训练提供丰富的样本空间。

这类算法模型可以提取抗原蛋白的空间特征、残基相对溶剂可及性、残基侧链属性、分子对接能量等特征，让模型学习到更多标签信息和特征差异，从而预测新的蛋白质结构中可能的B细胞抗原表位。Sun等[21]利用“残基三角形的单位斑块”的概念，提出了一种方法SEPPA来表征蛋白质表面的局部空间背景。SEPPA 2.0 [22]通过人工神经网络（ANNs）算法巩固了AAindex（氨基酸指数）的特征。SEPPA 3.0 [23]还进一步扩充了糖基化三角形和糖基化相关的AAindex。除了糖蛋白之外，其他翻译后修饰蛋白例如脂蛋白和金属蛋白也会影响表位的表达，目前仅有SEPPA3.0考虑翻译后修饰的糖蛋白表位预测，而其他相关蛋白质的预测算法也亟待研发。表2总结了进一步的基于结构的预测方法。

**Table 2 Lists the structure-based B-cell epitope prediction tools**

**表2 基于结构的B细胞抗原表位预测工具列表**

| 工具名称 | 发表年份 | 训练集/测试集 | 特征 | 算法 | ROC-AUC（文献中） |
|---|---|---|---|---|---|
| PEPITO [24] | 2008 | NA/215 | 基于氨基酸倾向评分和多距离半球暴露值的方法，线性组合计算表位分数 | NA | 0.754 |
| SEPPA [21] | 2009 | 82/119个表位 | 首次提出“残基三角形的单位斑块”的概念，利用氨基酸的倾向性索引和聚类系数来计算每个残基的预测得分 | NA | 0.742 |
| DiscoTope-2.0 [25] | 2012 | 75/NA | 采用了新的空间邻域定义，使用半球曝露作为表位测量方法 | NA | 0.727 |
| SEPPA-2.0 [22] | 2014 | 314/42 | 考虑了蛋白质抗原的亚细胞定位和免疫宿主的物种。引入了相对可及表面积（ASA）倾向性和综合氨基酸指数作为新的分类参数 | 逻辑回归 | 0.745~0.823 |

续表2

| 工具名称 | 发表年份 | 训练集/测试集 | 特征 | 算法 | ROC-AUC（文献中） |
|---|---|---|---|---|---|
| SEPPA-3.0[23] | 2019 | 767/236 | 针对糖蛋白抗原的空间表位预测进行了增强。引入了糖基化三角形和糖基化相关氨基酸指数作为新的分类器 | 逻辑回归 | 0.749~0.79 |
| epitope3D[26] | 2022 | 180/65 | 开发基于图特征的概念来建模区分表位和非表位区域 | AdaBoost | 0.78 |
| SEMA[27] | 2022 | 783/101 | 预测基于抗原的一级序列和三级结构的构象B细胞表位。使用深度学习技术改善B细胞表位的预测精度。提供了一个可解释的得分，表明与目标抗体的预期接触数 | 蛋白质语言模型ESM-1v和反向折叠模型ESM-IF1 | 0.76 |
| DiscoTope-3.0[28] | 2023 | 582/24 | 利用从解析结构和预测结构生成的逆折叠表示（使用ESM-IF1模型）。能够处理解析和预测的蛋白质结构，扩展了工具的适用范围 | XGBoost模型和反向折叠模型ESM-IF1 | 0.799 |

### 2.3　抗体特异性的B细胞抗原表位预测工具

上文已经提到了很多预测空间表位的方法，但它们只关注抗原，而忽略了同源抗体的信息。换句话说，这些方法预测的是抗原表面的所有表位残基，而这些抗原残基可能是多个抗体群的靶标，而不是特定的单克隆抗体。对于上文中基于结构的预测工具，由于缺少同源抗体的信息，这些算法计算的结果实际上是一种泛抗原表位位点。

如果研究人员想要探究抗体特异性的抗原表位，传统方法是将抗原-抗体相互作用视为一般的蛋白质-蛋白质相互作用，采用基于分子对接的策略。典型的方法包括ZDOCK[29]和ClusPro[30]，它们从形状、电子静力学和评分统计潜力等方面计算生物大分子之间的互补性。但是这种基于分子对接的方法得到的结果准确度并不理想，并且需要消耗大量的计算时间和资源。

所以为了探究针对特定抗体的抗原表位，实验人员开发了抗体特异性的B细胞抗原表位预测工具。EpiPred[31]提出了一种基于全局对接的算法来识别表位区域。Qiu等[32]提出了一种基于分子指纹的斑块模型SEPPA-mAb，对表位斑块和互补决定区（CDR）斑块之间的潜在互补性进行评分，预测抗体特异性表位。Desta等[33]提出一种基于ClusPro分子对接工具的抗体特异性抗原表位残基预测打分系统AbEmap。DeepMind最新发布的AlphaFold3[34]能够预测所有生命分子（蛋白质、DNA、RNA、配体等）的结构和相互作用，AlphaFold3在其他生命分子预测上效果很好，局部距离差测试（local distance difference test，LDDT）最佳能达到0.8，而在抗体抗原的相互作用上会明显低于蛋白质-蛋白质相互作用的性能，LDDT仅能达到0.4左右，且需要多次尝试才能达到最佳效果，这意味着抗原抗体是一类特殊的蛋白质-蛋白质相互作用，AlphaFold3虽然能够进行这方面的预测，还是需要设计专门针对抗原抗体相互作用的预测算法。表3展示了目前发表的抗体特异性B细胞表位预测工具。

**Table 3　List of antibody−specific B−cell epitope prediction tools**
**表3　抗体特异性的B细胞抗原表位预测工具列表**

| 工具名称 | 发表年份 | 训练集/测试集 | 特征 | 算法 | ROC-AUC（文献中） |
|---|---|---|---|---|---|
| EpiPred[31] | 2014 | 148/45 | 针对特定抗体的表位预测，可以整合到全局对接流程中，改进刚体对接算法的结果 | 几何拟合和基于知识的非对称抗体-抗原评分 | NA |
| SEPPA-mAb[32] | 2023 | 860/193 | 在SEPPA-3.0预测结果的基础上，基于指纹的斑块模型则对表位斑块和互补决定区（CDR）斑块之间的潜在互补性进行评分 | XGBoost | 0.774 |
| AbEmap[33] | 2023 | 40/21 | 支持从抗体的X射线结构、同源模型或仅有的氨基酸序列开始预测表位，集成了模板建模方法和抗原-抗体接触预测 | 蛋白质对接（PIPER程序）和同源建模（MODELLER）技术 | 0.736 |
| AlphaFold3[34] | 2024 | NA/65 | 能够预测包括蛋白质、核酸、小分子、离子和修饰残基在内的复杂生物分子的结构。显著提高了蛋白质-配体相互作用、蛋白质-核酸相互作用以及抗体-抗原预测的准确性。使用生成性扩散方法，可以处理各种化学成分的复杂性 | Diffusion Module、Transformer | NA |

NA：未提及。

## 3 T细胞抗原表位预测工具

T细胞抗原表位预测工具与B细胞抗原表位预测工具的预测目标不同，T细胞抗原表位是T细胞识别和结合的特定肽段，与MHC分子结合，从而被T细胞识别和激活。该类工具通过分析抗原表位肽片段，来预测它们是否能够被特定MHC结合提呈或者被特定的TCR识别并与其结合[35]。这些工具的设计旨在提供一种有效的方式，以降低实验室中对T细胞表位的识别和验证的成本和时间，从而帮助研究人员确定候选抗原表位肽，进而加速新药物研发过程。

### 3.1 T细胞表位肽与MHC结合提呈预测工具

目前已有多种预测工具被开发出来，用于预测T细胞表位肽与MHC分子的结合及提呈。这些工具使用不同的算法和数据来源，包括机器学习、深度学习、统计学等方法。

由丹麦技术大学（DTU）生物信息学中心开发的NetMHC系列[36]，是目前最为广泛使用的MHC与肽结合预测工具之一。该系列工具包括NetMHC、NetMHCpan、NetMHCII和NetMHCIIpan[37]，分别用于预测MHC I类和MHC II类分子的肽结合。O'Donnell等[38]开发了MHCflurry，使用机器学习模型预测MHC I类分子与肽的结合亲和力。它提供了一个用户友好的界面和应用程序编程接口（application programming interface，API），能够快速高效地进行大规模预测。MHCflurry通过使用大规模的实验数据训练模型，提供了高精度的预测结果。Shao等[39]使用深度学习模型来预测MHC I类和MHC II类分子与肽的结合，开发了MHCnuggets。深度学习在处理复杂数据和捕捉非线性关系方面表现出色，使得MHCnuggets能够提供高准确性的预测。该工具也具有开源和易于使用的特点。表4展示了目前发表的部分T细胞表位肽与MHC结合提呈预测工具。

**Table 4 List of predictive tools for T cell epitope peptide binding to MHC**

**表4 T细胞表位肽与MHC结合提呈预测工具列表**

| 工具名称 | 发表年份 | 训练集/测试集 | 特征 | 算法 | ROC-AUC（文献中） |
|---|---|---|---|---|---|
| NetMHC-3.0[36] | 2008 | 6 452/3 104 | 预测人类、小鼠和猴子的MHC I类分子对长度为8至11的肽段的亲和力 | 人工神经网络（ANN）和特定位置评分矩阵（PSSM） | 0.86 |
| NetMHCpan-4.1[37] | 2020 | 13 245 212个数据点/NA | 提供MHC I类分子的肽结合预测，结合了质谱（MS）洗脱配体数据和其他传统肽-MHC结合数据以提高预测性能 | NNAlign MA机器学习框架 | 0.95 |
| NetMHCIIpan[37] | 2020 | 4 086 230个数据点/NA | 提供MHC II类分子的肽结合预测，针对多种动物（包括人类、小鼠、牛、猿类、猪、马和狗）的MHC分子提供预测 | NNAlign MA机器学习框架 | 0.89 |
| MHCflurry[38] | 2020 | 493 473个MS（质谱）数据点和219 596个亲和力测量数据点/NA | 整合了MHC I类结合亲和力（BA）预测和抗原处理（AP）的预测。能够独立地预测MHC等位基因的效应和非等位基因依赖的效应 | 使用神经网络训练了两个独立的模型——MHC I类分子BA预测器和AP（抗原处理）预测器，并通过逻辑回归模型整合这两个模型的输出 | 0.91 |
| MHCnuggets[39] | 2020 | 241 553个肽-等位基因对的化学结合亲和力数据和96 211个肽-等位基因对的数据/26 888个$IC_{50}$测量数据 | 预测MHC I类和II类等位基因的肽结合。利用LSTM神经网络处理变长的肽输入。支持稀有等位基因的预测。可以整合结合亲和力和质谱HLAp数据来训练 | 长短期记忆（LSTM）神经网络 | 0.924 |

### 3.2 T细胞表位肽与TCR结合能力预测工具

预测T细胞表位肽与TCR结合能力的工具可以根据给定的抗原多肽序列和相关信息，预测该肽段是否与特定TCR结合。本质上，这个任务也是一个分类问题，模型需要预测结合与否的标签[40]。

主流的模型包括基于机器学习和深度学习的方

法。机器学习方法通常使用手工设计的特征，如序列相似性、物理化学性质等，然后利用这些特征训练分类器。而深度学习方法则可以端到端地学习从序列到结合预测的映射关系。模型根据给定的肽段和TCR序列，预测它们是否能够结合。这种预测能力对于理解T细胞免疫应答的机制、研究自身免疫性疾病以及设计个性化免疫疗法都具有重要意义。表5展示了目前发表的部分T细胞表位肽与TCR结合能力预测工具。

**Table 5　List of prediction tools for T cell epitope peptide binding ability to TCR**

**表5　T细胞表位肽与TCR结合能力预测工具列表**

| 工具名称 | 发表年份 | 训练集/测试集 | 特征 | 算法 | ROC-AUC（文献中） |
|---|---|---|---|---|---|
| NetTCR 2.0 [41] | 2021 | 9 204个CDR3β序列和2 744个配对肽数据点/NA | 预测TCR与MHC-肽复合体的结合。使用单链和双链数据来训练模型 | 一维卷积神经网络（1D CNN） | 0.89 |
| ERGO2 [42] | 2021 | McPAS和 VDJdb所有配对数据/NA | 集成了TCR的α和β CDR3序列，V和J基因，以及T细胞类型（CD4+或CD8+） | 长短期记忆（LSTM）网络和自编码器，多层感知机（MLP） | 0.7 |
| ImRex [43] | 2021 | 19 842/4 101 | 将CDR3和表位序列的物理化学属性转换为交互地图，这种方法避免了为每个分子独立创建内部嵌入 | 卷积神经网络（CNN） | 0.68 |
| pMTnet [44] | 2021 | 32 607/619 | 预测TCR与pMHC之间的结合特异性。应用迁移学习来利用大量TCR和pMHC数据 | 长短时记忆网络（LSTM）和堆叠自编码器 | 0.827 |
| Panpep [45] | 2023 | 23 232/5 230 | 设计用于识别TCR与各种类型肽的绑定，支持少数样本学习（few-shot）、零样本学习（zero-shot）和大多数样本学习（majority） | 元学习和神经图灵机 | 0.734 |
| TEIM [46] | 2023 | 45 481/122 | 预测CDR3β和表位之间所有残基对的距离和接触概率。应用少样本学习策略，使用大量序列级二元绑定数据进行预训练 | 二维卷积神经网络（2D CNNs） | 0.78 |
| DeepAIR [47] | 2023 | 38 558/6 423 | 利用深度学习框架，结合序列信息和结构信息，使用门控基于注意力的机制和张量融合技术进行特征整合，并通过特定任务的预测层进行AIR-抗原结合反应的预测 | 门控基于注意力的机制（gating-based attention mechanism）和张量融合机制（tensor fusion mechanism） | 0.912 |

NA：未提及。

## 4　机器学习算法在表位预测领域的应用

近年来，随着机器学习算法不断发展，越来越多的新的模型和算法被应用到B细胞抗原表位预测领域，其中表现较为出色的算法主要是极限梯度提升[48]（eXtreme Gradient Boosting，XGBoost）和蛋白质语言模型ESM（evolutionary scale modeling）。

XGBoost利用梯度提升的原理，通过逐步添加弱学习器（通常是决策树）来纠正前一轮学习器的错误，通过结合多个弱学习器形成一个强大的集成模型。XGBoost利用梯度提升能够有效捕捉生物数据中的复杂模式，最小化特定的损失函数，迭代地改进预测能力，使得预测结果比传统方法更可靠。并且B细胞抗原表位预测通常涉及高维数据，包括氨基酸序列、结构特征和生化性质，XGBoost在处理高维数据（具有大量特征）方面表现出色，这对于B细胞抗原表位预测领域尤为重要。该算法通过有效的特征选择和重要性度量来管理大特征空间，使其能够专注于最相关的特征并减少过拟合。

ESM是一种基于深度学习的蛋白质序列分析模型，旨在通过捕捉蛋白质序列中的进化信息来提高生物信息学任务的预测精度。ESM模型利用了自注意力机制（self-attention）和转换器（transformer）架构，能够在大规模蛋白质序列数据上进行训练，捕捉到蛋白质序列中的复杂模式和

结构信息。

研究人员使用ESM模型在大规模蛋白质序列数据上进行了预训练，捕捉到了丰富的进化信息和序列模式。这些信息对表位预测非常重要，因为表位通常具有保守性和特定的序列特征。ESM模型能够利用这些信息，有效提高表位预测的准确性。并且B细胞抗原表位的识别依赖于复杂的序列和结构模式。ESM模型的自注意力机制使其能够处理这些复杂的模式，识别出潜在的表位区域。相比传统的序列分析方法，ESM模型在捕捉长距离依赖关系和序列-结构关系方面表现更好。

由于ESM模型可以在大规模进化数据上进行预训练，它还具有较强的跨物种泛化能力。这对于表位预测尤为重要，因为疫苗和治疗性抗体开发需要考虑不同病原体和变种的表位识别能力。ESM模型能够在不同物种间识别保守的表位区域，有效提高预测的可靠性。

## 5 总结与展望

抗原表位的研究在揭示免疫系统对抗原的识别和应答机制方面发挥着至关重要的作用。B细胞和T细胞表位的识别与预测工具的不断发展为免疫学、疫苗设计以及个性化医疗领域提供了强大的支持。B细胞表位的预测工具包括基于序列、基于结构和抗体特异性3大类。3类工具通过分析蛋白质一维序列和三维结构数据提高预测的精确性。这些工具对于疫苗设计尤为重要，能够帮助科学家更好地了解抗原表位的三维结构，从而有针对性地设计疫苗，引导免疫系统产生更有效的抗体反应。T细胞表位的预测工具主要关注抗原肽与MHC结合以及pMHC与TCR结合两类。工具如NetMHCpan[37]等通过机器学习算法实现了对肽段与MHC分子结合的泛特异性预测，为深入理解免疫应答机制提供了有力工具。然而，面对MHC分子的巨大多样性和预测模型的挑战，预测的准确性仍然是一个亟待解决的问题。

未来，抗原表位预测领域面临着巨大的机遇和挑战。随着计算生物学、生物信息学和机器学习领域的不断进步，我们有望更全面、准确地预测抗原表位。基于深度学习的方法，如SEMA[27]和DiscoTope[28]，以及基于图特征的Epitope3D[26]等工具的出现，为未来研究提供了新的思路和方法。这些工具的应用将不仅加深我们对抗原表位的理解，也推动医学研究在疾病治疗、个性化医疗和疫苗设计等方面取得更大的突破。此外，抗体特异性B细胞表位预测工具的涌现，如SEPPA-mAb[32]和AbEmap[33]，为抗体工程和治疗提供了新的方向。这些工具的应用将有望加速单克隆抗体的研发过程，为临床治疗提供更加个性化和有效的解决方案。在T细胞表位预测方面，泛特异性预测方法的不断创新，如Panpep[45]、MHCnuggets[39]和NetMHCpan[37]等，为未知MHC分子的预测提供了新的思路。然而，面对免疫系统中MHC分子的巨大复杂性，未来的研究需要更深入地挖掘MHC与抗原肽结合的机制，以更全面地理解T细胞抗原识别的规律。总之，抗原表位预测领域的进展将进一步推动免疫学、生物医学和药物研究的发展。通过不断优化和创新预测工具，我们有望更好地理解免疫系统对抗原的响应，为未来个性化医学和精准治疗提供更为可靠的科学依据。在这个充满挑战和机遇的领域，科学家们将继续努力，推动抗原表位预测技术走向新的高峰。

当前，抗原表位预测技术已成为免疫学研究和疫苗设计中不可或缺的工具，尤其是在精确识别B细胞和T细胞表位方面。尽管已取得了显著进展，但该领域仍面临若干挑战和提升空间，未来的发展方向也显得尤为重要。目前的预测工具虽然能够处理复杂数据，但预测准确性和效率仍有待提高。通过整合更多的生物信息学数据、采用更先进的机器学习算法，如深度学习和迁移学习，有望进一步优化预测性能。当然，研究人员还可以提高工具的泛化能力，使其能够跨物种预测抗原表位，对于理解宿主跨物种的免疫反应和开发广谱疫苗具有重要意义。同时，抗体特异性表位预测工具近年来也成为了研究热点，结合抗原的三维结构信息和功能数据，发展能够同时考虑抗原结构和抗体结合动态性的预测工具将成为一个重要的研究方向。这样的工具可以更好地模拟抗原-抗体之间的真实相互作用，对于抗体药物的开发、个性化医疗以及疫苗设计尤其有价值。随着大预言模型的发展，加强预测工具的可解释性越来越重要，这可以帮助研究者更好地理解预测结果背后的生物学意义。同时，开发更加用户友好的界面，降低使用门槛，促进工具的广泛应用。

## 参 考 文 献

[1] Korber B, LaBute M, Yusim K. Immunoinformatics comes of age. PLoS Comput Biol, 2006, **2**(6): e71

[2] Bonilla F A, Oettgen H C. Adaptive immunity. J Allergy Clin Immunol, 2010, **125**(2 suppl 2): S33-S40

[3] Oli A N, Obialor W O, Ifeanyichukwu M O, *et al*. Immunoinformatics and vaccine development: an overview. Immunotargets Ther, 2020, **9**: 13-30

[4] Burger J A, Wiestner A. Targeting B cell receptor signalling in cancer: preclinical and clinical advances. Nat Rev Cancer, 2018, **18**(3): 148-167

[5] Palatnik-de-Sousa C B, Soares I S, Rosa D S. Editorial: epitope discovery and synthetic vaccine design. Front Immunol, 2018, **9**: 826

[6] Zeng X, Bai G, Sun C, *et al*. Recent progress in antibody epitope prediction. Antibodies (Basel), 2023, **12**(3): 52

[7] Kumar N, Bajiya N, Patiyal S, *et al*. Multi-perspectives and challenges in identifying B-cell epitopes. Protein Sci, 2023, **32**(11): e4785

[8] Farrera-Soler L, Daguer J P, Barluenga S, *et al*. Experimental identification of immuno- dominant B-cell epitopes from SARS-CoV-2. Chimia, 2021, **75**(4): 276-284

[9] Accardi A, Lobet S, Williams C, *et al*. Synergism between halide binding and proton transport in a CLC-type exchanger. J Mol Biol, 2006, **362**(4): 691-699

[10] Peters B, Nielsen M, Sette A. T cell epitope predictions. Annu Rev Immunol, 2020, **38**: 123-145

[11] Schaap-Johansen A L, Vujović M, Borch A, *et al*. T cell epitope prediction and its application to immunotherapy. Front Immunol, 2021, **12**: 712488

[12] Cole D K, Bulek A M, Dolton G, *et al*. Hotspot autoimmune T cell receptor binding underlies pathogen and insulin peptide cross-reactivity. J Clin Invest, 2016, **126**(6): 2191-2204

[13] Galanis K A, Nastou K C, Papandreou N C, *et al*. Linear B-cell epitope prediction for *in silico* vaccine design: a performance review of methods available *via* command-line interface. Int J Mol Sci, 2021, **22**(6): 3210

[14] Cia G, Pucci F, Rooman M. Critical review of conformational B-cell epitope prediction methods. Brief Bioinform, 2023, **24**(1): bbac567

[15] Rubinstein N D, Mayrose I, Martz E, *et al*. Epitopia: a web-server for predicting B-cell epitopes. BMC Bioinformatics, 2009, **10**: 287

[16] Ansari H R, Raghava G P. Identification of conformational B-cell Epitopes in an antigen from its primary sequence. Immunome Res, 2010, **6**: 6

[17] Jespersen M C, Peters B, Nielsen M, *et al*. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res, 2017, **45**(W1): W24-W29

[18] Clifford J N, Høie M H, Deleuran S, *et al*. BepiPred-3.0: improved B-cell epitope prediction using protein language models. Protein Sci, 2022, **31**(12): e4497

[19] da Silva B M, Ascher D B, Pires D E V. epitope1D: accurate taxonomy-aware B-cell linear epitope prediction. Brief Bioinform, 2023, **24**(3): bbad114

[20] Berman H M, Westbrook J, Feng Z, *et al*. The protein data bank. Nucleic Acids Res, 2000, **28**(1): 235-242

[21] Sun J, Wu D, Xu T, *et al*. SEPPA: a computational server for spatial epitope prediction of protein antigens. Nucleic Acids Res, 2009, **37**(Web Server issue): W612-W616

[22] Qi T, Qiu T, Zhang Q, *et al*. SEPPA 2.0—more refined server to predict spatial epitope considering species of immune host and subcellular localization of protein antigen. Nucleic Acids Res, 2014, **42**(Web Server issue): W59-W63

[23] Zhou C, Chen Z, Zhang L, *et al*. SEPPA 3.0-enhanced spatial epitope prediction enabling glycoprotein antigens. Nucleic Acids Res, 2019, **47**(W1): W388-W394

[24] Sweredoski M J, Baldi P. PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. Bioinformatics, 2008, **24**(12): 1459-1460

[25] Kringelum J V, Lundegaard C, Lund O, *et al*. Reliable B cell epitope predictions: impacts of method development and improved benchmarking. PLoS Comput Biol, 2012, **8**(12): e1002829

[26] da Silva B M, Myung Y, Ascher D B, *et al*. epitope3D: a machine learning method for conformational B-cell epitope prediction. Brief Bioinform, 2022, **23**(1): bbab423

[27] Shashkova T I, Umerenkov D, Salnikov M, *et al*. SEMA: antigen B-cell conformational epitope prediction using deep transfer learning. Front Immunol, 2022, **13**: 960985

[28] Høie M H, Gade F S, Johansen J M, *et al*. DiscoTope-3.0: improved B-cell epitope prediction using inverse folding latent representations. Front Immunol, 2024, **15**: 1322712

[29] Pierce B G, Wiehe K, Hwang H, *et al*. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics, 2014, **30**(12): 1771-1773

[30] Kozakov D, Hall D R, Xia B, *et al*. The ClusPro web server for protein-protein docking. Nat Protoc, 2017, **12**(2): 255-278

[31] Krawczyk K, Liu X, Baker T, *et al*. Improving B-cell epitope prediction and its application to global antibody-antigen docking. Bioinformatics, 2014, **30**(16): 2288-2294

[32] Qiu T, Zhang L, Chen Z, *et al*. SEPPA-MAb: spatial epitope prediction of protein antigens for mAbs. Nucleic Acids Res, 2023, **51**(W1): W528-W534

[33] Desta I T, Kotelnikov S, Jones G, *et al*. The ClusPro AbEMap web server for the prediction of antibody epitopes. Nat Protoc, 2023, **18**(6): 1814-1840

[34] Abramson J, Adler J, Dunger J, *et al*. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 2024, **630**(8016): 493-500

[35] Lybaert L, Lefever S, Fant B, *et al*. Challenges in neoantigen-directed therapeutics. Cancer Cell, 2023, **41**(1): 15-40

[36] Lundegaard C, Lamberth K, Harndahl M, *et al*. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. Nucleic Acids Res, 2008, **36**(Web Server issue): W509-W512

[37] Reynisson B, Alvarez B, Paul S, *et al*. NetMHCpan-4.1 and

NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res, 2020, **48**(W1): W449-W454

[38] O'Donnell T J, Rubinsteyn A, Laserson U. MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. Cell Syst, 2020, **11**(1): 42-48.e7

[39] Shao X M, Bhattacharya R, Huang J, *et al*. High-throughput prediction of MHC class I and II neoantigens with MHCnuggets. Cancer Immunol Res, 2020, **8**(3): 396-408

[40] Desai D V, Kulkarni-Kale U. T-cell epitope prediction methods: an overview. Methods Mol Biol, 2014, **1184**: 333-364

[41] Montemurro A, Jessen L E, Nielsen M. NetTCR-2.1: lessons and guidance on how to develop models for TCR specificity predictions. Front Immunol, 2022, **13**: 1055151

[42] Springer I, Tickotsky N, Louzoun Y. Contribution of T cell receptor alpha and beta *CDR3 MHC* typing, V and J genes to peptide binding prediction. Front Immunol, 2021, **12**: 664514

[43] Moris P, de Pauw J, Postovskaya A, *et al*. Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. Brief Bioinform, 2021, **22**(4): bbaa318

[44] Lu T, Zhang Z, Zhu J, *et al*. Deep learning-based prediction of the T cell receptor-antigen binding specificity. Nat Mach Intell, 2021, **3**(10): 864-875

[45] Gao Y, Gao Y, Fan Y, *et al*. Pan-Peptide Meta Learning for T-cell receptor-antigen binding recognition. Nat Mach Intell, 2023, **5**: 236-249

[46] Peng X, Lei Y, Feng P, *et al*. Characterizing the interaction conformation between T-cell receptors and epitopes with deep learning. Nat Mach Intell, 2023, **5**: 395-407

[47] Zhao Y, He B, Xu F*et al*. DeepAIR: a deep learning framework for effective integration of sequence and 3D structure to enable adaptive immune receptor analysis. Sci. Adv, 2023, **9**(32): eabo5128

[48] Chen T Q, Guestrin C. XGBoost: a scalable tree boosting system. arXiv, 2016. DOI: 10.48550/arXiv.1603.02754

# Current Research and Development of Antigenic Epitope Prediction Tools*

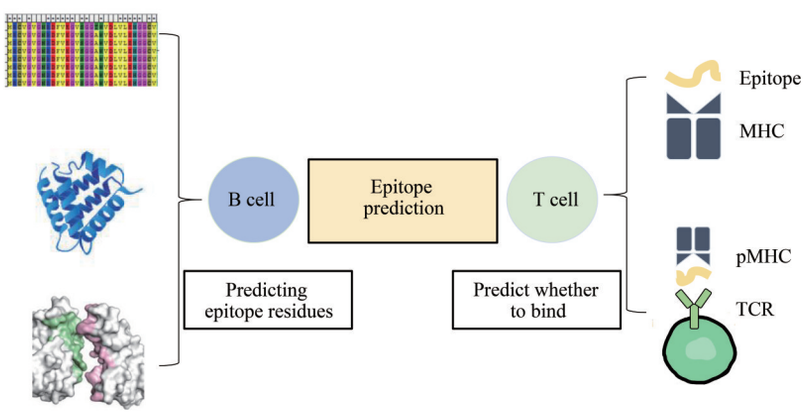
LI Zi-Hao[1)], WANG Yuan[2)], MAO Tian-Tian[2)], CAO Zhi-Wei[1)**], QIU Tian-Yi[3)**]

([1)]*School of Life Sciences, Fudan University, Shanghai* 200433, China;

[2)]*College of Life Science and Technology, Tongji University, Shanghai* 200092, China;

[3)]*Zhongshan Hospital Experimental Research Center, Fudan University, Shanghai* 200032, China)

## Graphical abstract




**Abstract**　Adaptive immunity is a critical component of the human immune system, playing an essential role in identifying antigens and orchestrating a tailored immune response. This review delves into the significant strides made in the development of epitope prediction tools, their integration into vaccine design, and their pivotal role in enhancing immunotherapy strategies. The review emphasizes the transformative potential of these tools in refining our understanding and application of immune responses. Adaptive immunity distinguishes itself from innate immunity by its ability to recognize specific antigens and remember past infections, leading to quicker and more effective responses upon subsequent exposures. This facet of immunity involves complex interactions between various cell types, primarily B cells and T cells, which recognize distinct epitopes presented by antigens. Epitopes are small sequences or configurations on antigens that are recognized by the immune receptors on B cells and T cells, acting as the focal points of immune recognition and response. Epitopes can be broadly classified into two types: linear (or sequential) epitopes and conformational (or discontinuous) epitopes. Linear epitopes consist of a sequence of amino acids in a protein that are recognized by B cells and T cells in their primary structure form. Conformational epitopes, on the other hand, are formed by spatially distinct amino acids that come together in the tertiary structure of the protein, often recognized by the immune system only when the protein folds into its native conformation. The role of epitopes in the immune response is critical as they are the

* This work was supported by a grant from The National Natural Science Foundation of China（32370697）.

** Corresponding author.

CAO Zhi-Wei. Tel：86-21-65980296，E-mail：zwcao@fudan.edu.cn

QIU Tian-Yi. Tel：86-21-64041990，E-mail：tianyi_qiu@fudan.edu.cn

Received：July 30，2024　Accepted：September 3，2024

primary triggers for the activation of B cells and T cells. When an epitope is recognized, it can stimulate B cells to produce antibodies, mobilize helper T cells to secrete cytokines, or prompt cytotoxic T cells to kill infected cells. These actions form the basis of the adaptive immune response, tailored to eliminate specific pathogens or infected cells effectively. The prediction of B cell and T cell epitopes has evolved with advances in computational biology, leading to the development of several sophisticated tools that utilize a variety of algorithms to predict the likelihood of epitope regions on antigens. Tools employing machine learning methods, such as support vector machines (SVMs), XGBoost, random forest, analyze large datasets of known epitopes to classify new sequences as potential epitopes based on their similarity to known data. Moreover, deep learning has emerged as a powerful method in epitope prediction, leveraging neural networks capable of learning high-dimensional data from vast amounts of immunological inputs to identify patterns that may not be evident to other predictive models. Deep learning models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs) and ESM protein language model have demonstrated superior accuracy in mapping the nonlinear relationships inherent in protein structures and epitope interactions. The application of epitope prediction tools in vaccine design is transformative, enabling the development of epitope-based vaccines that can elicit targeted immune responses against specific parts of the pathogen. These vaccines, by focusing the immune response on highly specific regions of the pathogen, can offer high efficacy and reduced side effects. Similarly, in cancer immunotherapy, epitope prediction tools help identify tumor-specific antigens that can be targeted to develop personalized immunotherapeutic strategies, thereby enhancing the precision of cancer treatments. The future of epitope prediction technology appears promising, with ongoing advancements anticipated to enhance the precision and efficiency of these tools further. The integration of broader immunological data, such as patient-specific immune profiles and pathogen variability, along with advances in AI and machine learning, will likely drive the development of more adaptive, robust, and clinically relevant prediction models. This will not only improve the effectiveness of vaccines and immunotherapies but also contribute to our broader understanding of immune mechanisms, potentially leading to breakthroughs in the treatment and prevention of multiple diseases. In conclusion, the development and refinement of epitope prediction tools stand as a cornerstone in the advancement of immunological research and therapeutic design, highlighting a path toward more precise and personalized medicine. The ongoing integration of computational models with experimental immunology holds the promise of revolutionizing our approach to combating infectious diseases and cancer.

**Key words** antigen epitope prediction, adaptive immunity, B cells, T cells, machine learning, deep learning
**DOI**：10.16476/j.pibb.2024.0351