

# 目录

<b>第一章 研究背景</b>	<b>1</b>
第 1 节 抗原抗体的相互作用	1
第 2 节 抗原抗体相互作用研究现状	5
第 3 节 亲和力预测	13
第 4 节 本论文的主要内容	14
<b>第二章 数据的提取和初步分析</b>	<b>15</b>
第 1 节 数据提取	15
第 2 节 数据分析	19
第 3 节 节标题	19
第 4 节 子节标题	20
第 5 节 正文	20
第 6 节 章节	20
<b>第三章 公式排版</b>	<b>21</b>
第 1 节 行内公式	21
第 2 节 行间公式	21
<b>第四章 表格和图片</b>	<b>23</b>
<b>第五章 定理环境</b>	<b>24</b>
第 1 节 题头	24
第 2 节 同章另一节的题头	24
<b>第六章 参考文献的写法</b>	<b>25</b>

# 数学学院毕业论文模版

刘传省

学号：17210180030

专业：计算系统生物学

**摘要** 这是我的中文摘要

**关键字：**正文写法, 公式写法, 参考文献写法.

**Abstract** This is my English abstract.

**Keywords:** 正文写法, 公式写法, 参考文献写法.

# 第一章 研究背景

本章主要介绍相关研究的背景知识，要解决问题的意义和研究进展

## 第1节 抗原抗体的相互作用

人的免疫系统是人体抵抗外界病原入侵的重要系统，它可以分为天然免疫(innate immunity)和获得性免疫(adaptive immunity)。天然免疫不具有特异性，或者最多也只能针对一大类的病原进行防御。它包括巨噬细胞、抑菌蛋白、NK细胞、补体系统、粒细胞等。天然免疫构成了人体防御的第一道防线。一旦病原突破第一道防线，人体就要进行获得性免疫。获得性免疫是针对入侵的病原产生一系列的特异性的免疫反应，包括特异性的细胞免疫和体液免疫。获得性免疫的特异性，可以使得人体把主要的资源集中起来应对特定的病原，从而更高效。但是，自然界的病原千千万万，那么获得性免疫是如何识别这不同的病原的呢？

对于特异性的细胞免疫来讲，他的特异性可以用下面的图 1.1 来说明。

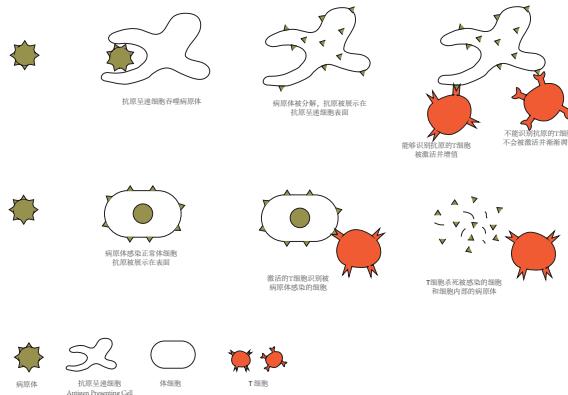


图 1.1: 特异性细胞免疫示意图。此图知识简要说明了细胞免疫的过程，真是的情况比这里要复杂的多，比如共刺激等。

对不同的病原体来讲，都有其独特的结构和成分，那些可以引起免疫反应的结构和成分，成为抗原(antigen)。体液免疫的特异性就是特定抗体对特定抗原的识别，如图1.2。

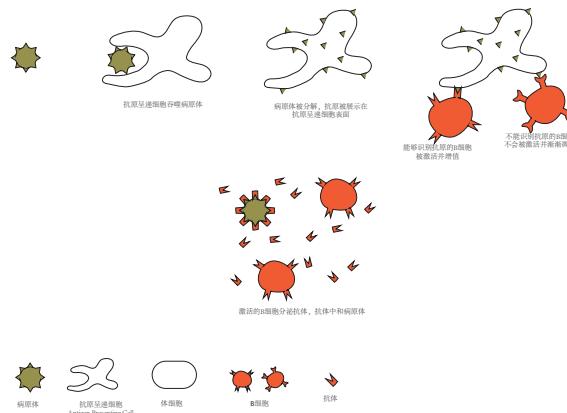


图 1.2: 特异性体液免疫示意图。此图知识简要说明了体液免疫过程中，抗体的产生和对病原的识别，真实的过程比这要复杂的多，比如说 T helper 的作用等。同时，抗体除了直接杀死病原体之外，还可以参与抗体介导的细胞毒性(antibody directed cell cytotoxicity)。

人类的抗体结构是一个二聚体，由两条重链(heavy chain)和两条轻链(light chain)组成。每条链又分为可变区(variable fragment)和不变区(constant fragment)。抗体的特异性则主要来自与可变区的互补决定区(CDR, complementarity determining region)如图1.3。

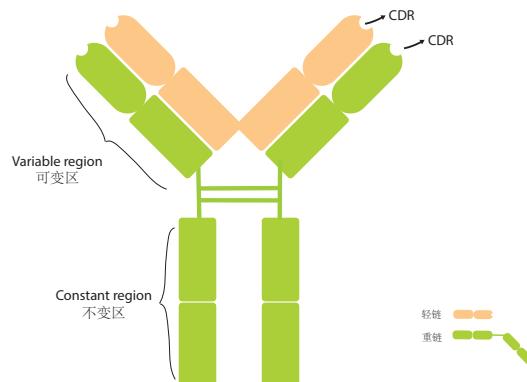


图 1.3: 抗体结构示意图

互补决定区主要由6个比较短的氨基酸片段组成，它们分为来自重链和轻链的CDR1, CDR2和CDR3。氨基酸在这些区域上的不同序列决定了抗体的特异性和多样性。对于抗原和抗体相互作用的研究，可以在一定程度上简化为CDRs和抗原局部区域的相互作用。

用。抗原上那些和抗体相互作用的部分又称为抗原表位(epitope)。CDR多样性的来源主要有两个，一个是不同基因片段的拼接，另外一个是细胞超突变(Somatic hypermutation)，也就是这些区域比其他区域有更高的突变率，有时候还会在拼接的过程中加入或者丢失一些碱基。理论上讲，可以产生的多样性可以达到 $10^{12}$ 数量级，甚至还要更多。所以，体液免疫是一个强大的免疫机制，几乎对所有抗原都可以产生对应特异性抗体。

体液免疫早在很久以前就被用来和疾病斗争。早在宋朝的时候，智慧的中国人就用“种痘”来预防天花，就是利用的用毒性弱的毒株让人体产生抗体和免疫记忆。这就是最早的疫苗了，只是这时种的是人痘，具有极高的风险。1796年，英国人Edward Jenner用接种牛痘的方法来预防天花，极大的降低了接种的风险。1979年，世界卫生组织(WHO,world health organization)宣布天花从地球上消除。这是，人类利用体液免疫的一次巨大成功，也是人类医学史上的壮举。自1796年以来，随着每次技术的进步，疫苗的数量和质量都会有很大的提升。

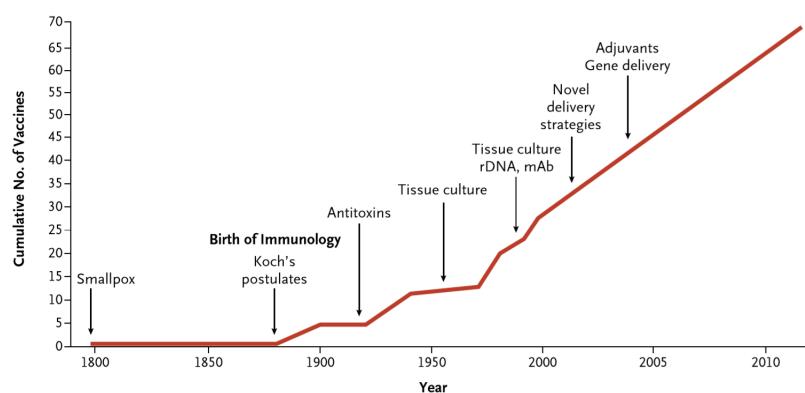


图 1.4: 疫苗的数量变化和疫苗开发技术的发展[1]

但是，并不是所有的传染病都能顺利开发出疫苗，比如说HIV-1[2]。其中的一个原因在于抗原的多变性。但是，也发现了一些具有广谱作用的抗体，可以抵抗多种不同的毒株。对这些抗体的进一步分析发现，它们对可以识别HIV-1一些保守的抗原表位(epitope)。知道了这些抗原表位之后，就可以通过抗原表位的嫁接(grafting)或者把抗原表位整合的特殊设计的架构(scaffolding)中，由此设计的疫苗会比传统意义上的疫苗效果更好。抗原表位的确定除了可以通过实验手段的到，还可以通过计算的手段，通过一系列的模型预测。

对抗原抗体相互作用的研究，除了可以帮助设计疫苗和预测抗原表位外，还可以促进对治疗性抗体(therapeutic antibody)的开发。随着单克隆抗体(monoclonal antibody)和人源化抗体(humanized antibody)技术的进步，越来越多的治疗性抗体被注册成新的药物。到2020年2月初，已经被FDA(Food and Drug Administration)和EMA(European

Medicines Agency)批准或正在审核的治疗性抗体就多达106个[3]。更是开发出很多具有广谱抗癌作用的治疗性抗体，其中有很多是针对免疫过程中的检测点(checkpoint)开发的[4]。比如，在治疗美国前总统吉米卡特的癌症中起着至关重要作用的抗体 pembrolizumab 就是通过抑制PD1(programmed cell death protein 1)，从而实现免疫细胞对癌细胞的杀伤。鉴于传统小分子药物开发越来越困难，以及抗体的多样性及相关技术的发展，治疗性抗体必将开辟人类药学史上一个新的时代[5]。

一个简单的单克隆抗体的生产过程大概如图[?]:

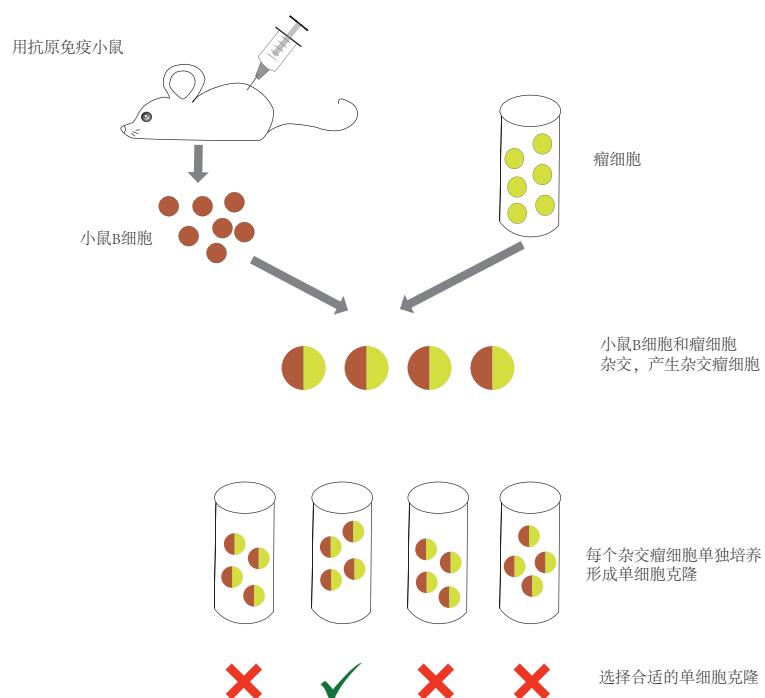


图 1.5: 单克隆抗体制备示意图

然而如果把这些由老鼠产生的单克隆抗体直接注射到人体内，往往会产生免疫反应。一个避免免疫反应的做法是把这些单克隆抗体的CDR区域的氨基酸序列安插到人类抗体对应的位置，这样的抗体就是人源化抗体。CDR序列的产生需要大量的实验，如果这个过程可以通过计算的手段来做比较准确的预测，则会对抗体的制备有深刻影响。同时，即便是通过实验手段产生CDR序列，由于实验过程中一些比较难以控制的因素，产生的序列也未必能满足我们的要求。一个不易控制的因素是抗体的作用位点。对于一个抗原来讲，可能的抗原表位会有很多，其中的任何一个抗原表位都可能诱导免疫反应，产生抗体，而理想的抗体往往需要针对特定的抗原表位。

另外一个不易控制的因素是抗原抗体之间的亲和力(affinity)。一个可以用作治疗

用的抗体往往要求具有足够高的亲和力,这直接关系到抗体的疗效。虽然免疫系统本身会筛选出亲和力比较高的抗体，但是无法保证这样的亲和力就满足要求。那么这就需要在原来抗体的基础上，对CDR序列进行一定的改造，从而提高亲和力，达到我们的要求。对抗原抗体的相互作用有足够的了解可以指导对这些序列的改造，从而大大节约抗体药物的研发和生产成本。

对抗原抗体相互作用的研究，除了上面说的意义之外，还有很多外溢效应。比如说设计合适的抗体来催化一些反应，也就是抗体酶。再比如说，设计一些治疗性的多肽。从更大的范围讲，抗原抗体的相互作用是蛋白与蛋白相互作用的一部分，研究蛋白和蛋白相互作用的方法，在一定程度上可以用来研究抗原抗体的相互作用。但是抗原抗体的相互作用又有其特殊性，因为抗原抗体的相互作用主要表现为抗体的CDR loop 区和抗原表位的相互作用。

## 第 2 节 抗原抗体相互作用研究现状

对抗原抗体相互作用，比较早的是Cothia，他第一次指出了抗体主要通过CDR区域和抗原相互作用，并且分析了CDR1和CDR2的经典结构。但是，这些都是描述性的，并不能对抗体的性质，以及什么样的抗原和什么样的抗体结合做出回答。接下来，大家开始对特定的抗原抗体复合物进行研究。其中对蛋清溶菌酶(hen egg white lysosome)和其抗体的相互作用的研究尤其多。Padlan 等解析了HyHEL-10 Fab和蛋清溶菌酶(HEL)的结构，认为抗原表位是不连续的，范德华力(van der Waals)和氢键(hydrogen bond)是相互作用的关键[6]；Yokota 等通过对HyHEL-10-HEL复合物中一些氨基酸的突变(L-Y50F, L-S91A, 和 L-S93A)来研究氢键的作用[7]，后来又研究了Arg的作用[8]；Pons 等通过Alanine scanning 研究了HyHEL-10-HEL中参与相互作用的各个氨基酸的重要性[9]；Shiroishi 等通过把Tyr 突变成 Phe 和 Ala 来研究 Tyr 在 HyHEL-10-HEL 中的作用[10]；同样，Shiroishi 等通过对 HyHEL-10-HEL 的分析，研究了盐桥(salt-bridge)的作用[11]；Kam-Morgan 等通过突变对HyHEL-10-HEL中抗原表位做了更为精细的研究[12]。除了HyHEL-10-HEL，还有许多文章对抗体HyHEL-63 和 HEL 的复合物HyHEL-63-HEL，以及HyHEL-5-HEL 做了许多类似的研究[13, 14, 15, 16, 17, 18, 19, 20, 21]。除了针对HEL 和其抗体的研究之外，还有许多关于其他抗原抗体的研究。但是，所有这些研究，都是关于某一个具体的复合物在特定情况下的氨基酸的作用，或者某种相互作用力的贡献，从来没有一个系统的关于所有抗原抗体相互作用的描述。其中一个主要原因，是抗原抗体相互作用的构象(conformation)性质。也就是说，参与抗原抗体相互作用的氨基酸，特别是抗原表位上的氨基酸，并不是线性排列的，而是具有空间上的关系。就拿1918年H1N1流感应大爆发时候流感病毒表面的血凝素(Hemagglutinin)SC1918/H1和它的抗体CR6261来说。

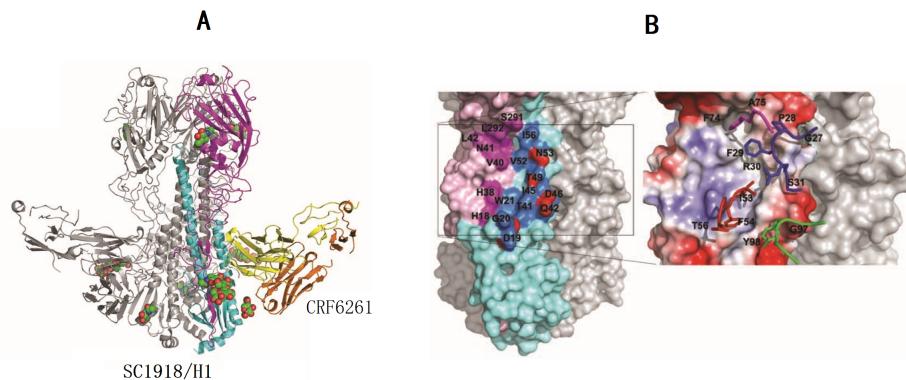


图 1.6: **A**是SC1918/H1 和 CR6261 复合物的结构。**B** 是 SC1918/H1 和抗体结合区域的放大图，其抗原表位都已经标出。此图由文献[22]中的图片编辑而成。

从图1.6 B中可以看出，在序列上，这些氨基酸并在序列不连续，然而在空间上却比较临近，形成有效的抗原表位，这就是抗原表位的构象性质。根据 Saba Ferdous 等最近对488个B-cell epitope 的研究，只有大概4%的是线性的。如果把有不小于3个的氨基酸参与相互作用，并且这些氨基酸在序列上的距离不大于3的位置定为一个区域(region)，那么只有约14%的抗原表位只有一个区域[23]，也就是说，抗原的表位是高度的非线性的。

虽然抗原表位的非线性特征，给抗原抗体的研究带了的巨大的困难。但是，鉴于抗原表位在抗体设计和疫苗研发中的重要性，对抗原表位预测的努力一直没有停止。从实验的角度来讲，主要有结构生物学的方法和突变的方法。结构生物学的方法是通过分析抗原抗体复合物的结构，来判断哪些是抗原表位。突变的方法则是通过在不同的位点引入突变来确定，究竟哪些氨基酸在抗原抗体的结合过程中起关键作用[24]。除了上面的实验手段，利用计算手段对抗原表位的预测，一直在发展。Rubinstein 等做了一个统计检验， $H_0$ 设定为抗原表位和非抗原表位没有差别， $H_a$ 设定为抗原表位和非抗原表位有差别。通过对大量抗原抗体的分析，认为抗原表位在氨基酸偏好、二级结构(secondary structure)、几何形状和进化的保守性上都和非抗原表位有显著区别[25]。虽然这篇文章发表在2008年，但是比这早的以及在这之后的许多抗原表位预测方法的理论依据，都是对任意的抗原来预测它的抗原表位。例如，Hopp等通计算局部氨基酸序列的亲水性(hydrophilicity)来预测抗原表位,认为亲水性最高的区域要么是抗原表位要么和抗原表位相邻[26]；Rubinstein等通过对一直的抗原表位的分析，用机器学习的方法来预测给定抗原的抗原表位；Jing Sun 等通过引入残基三角(residual triangle)的概念，计算每个氨基酸的倾向指数(propensity index)和集群系数(cluster coefficient)，由此来预测非线性抗原表位[28]；Xu等对 CEP[29],DiscoTope[30],PEPOP[31]、ElliPro[32]、BEpro[33] 和 SEPPA，六种预测方法在实验验证的数据集上进行了比较。

即便是最好的SEPPA，其AUC也只能达到0.62，原不满足实际需要[34]。

上面的方法都是对任意给定的抗原，来预测抗原表位，并不涉及到和抗原结合的抗体。理论上讲，一个抗原的任何区域都有可能被抗体识别，也就是任何区域都有可能是抗原表位，所以对一个抗原来讲，预测给定抗体的抗原表位才更有意义[35]。Shinji Sog 构建了ASEP 指数用于抗原表位，这是第一次在给定抗体情况下对抗原表位的预测[36]。具体的过程可以概括如下：

- (1) 搜集训练集。在PDB 数据库中搜集抗原抗体复合物，同时搜集其他同源二聚体(homodimer)或者异源二聚体(heterodimer)作为抗原抗体复合物的参照，这里为了方便，我们称为参照复合物。
- (2) 计算氨基酸 $x$ 在抗原表位中的出现频率 $f_e(x)$ :

$$f_e(x) = \frac{n_e(x)}{\sum_{y=1}^{20} n_e(y)}$$

其中， $n_e(x)$  是训练集中抗原抗体复合物的抗原表位中 $x$ 氨基酸的频率，分母中 $n_e(y)$ 做类似的解读，指标 $y$ 从1遍历到20是指对20中不同氨基酸的频率求和。

- (3) 计算氨基酸 $x$ 在参照复合物的表面氨基酸(surface residue)中的出现频率 $f_s(x)$ 。这里的表面氨基酸被作者定义为出现在复合物表面但是又不参与复合物的相互作用。

$$f_s(x) = \frac{n_s(x)}{\sum_{y=1}^{20} n_s(y)}$$

其中， $n_s(x)$  是参照复合物的表面氨基酸中 $x$ 氨基酸的频率，分母中 $n_s(y)$ 做类似的解读，指标 $y$ 从1遍历到20是指对20中不同氨基酸的频率求和。

- (4) 给定抗原中一个特定位置的抗原表位氨基酸 $x$ ，定义抗原抗体氨基酸对为 $xy$ ，其中 $y$ 为距离 $x$ 最近的位于抗体上的氨基酸。计算抗原抗体氨基酸对 $xy$ 的出现率 $f_{ep}(xy)$ :

$$f_{ep}(xy) = \frac{n_{ep}(xy)}{\sum_{z=1}^{20} n_{ep}(xz)} \times f_e(x)$$

- (5) 用传统的不依赖于抗原的方法DiscoTope 来预测给定抗原的抗原表位,这里我们简称为传统预测抗原表位。

- (6) 计算传统预测抗原表位中氨基酸 $x$ 的氨基酸频率 $f'_e(x)$ :

$$f'_e(x) = \sum_{y=1}^{20} [f_{ep}(xy) \times n_p(y)]$$

其中 $n_p(y)$ 是给定抗体的抗体表位(paratope)中氨基酸 $y$ 的出现频率。

- (7) 计算传统预测抗原表位中氨基酸 $x$ 抗体特意的抗原表位倾向ASEP(antibody-specific epitope propensity),  $p'_e(x)$ :

$$p'_e(x) = \log \left( \frac{f'_e(x)}{f_s(x)} \right)$$

- (8) 对于传统预测抗原表位中的氨基酸 $i$ , 计算:

$$\text{ASEP}(i) = \sum_{x=1}^{20} p'_e(x) c_i(x)$$

其中 $c_i(x)$ 是指在传统预测抗原表位中的氨基酸 $\alpha$ 碳与 $i$ 的 $\alpha$ 碳距离小于10Å的类型为 $x$ 的氨基酸的个数。这里也就是对 $i$ 附近的 $p'_e(x)$ 根据不同氨基酸的频率进行加权。

- (9) 对传统预测抗原表位中的每个氨基酸都计算ASEP, 然后排序, 取舍。ASEP越高越可能为抗原表位。

在上面的步骤中, (4)通过加权形式, 把抗原表位和抗体表位结合在一起, 从而在后面的计算过程中实现了在给定抗体情况下对抗原表位的ASEP打分。这个简单的加权, 比之前仅仅依靠抗原对抗原表位做出的预测有了明显提高。

Konrad Krawczy等在它们对蛋白与蛋白相互作用研究的基础上设计了在给定抗体情况下抗原表位的预测方法: EpiPred[39]。该方法在抗原表位预测的时候, 结合了抗原表位和抗体表位在空间上的互补性和一个由 i-Patch 转化而来的打分。一个氨基酸对的i-Patch 值的大概计算方法如下, 具体详细的计算请参考文献[40, 41]。

- (1) 为了减少可能出现的氨基酸对的数目, 把氨基酸进行分类, 文献[41]把氨基酸分成7类。

- (2) 计算两个氨基酸类( $C_1, C_2$ )之间的倾向性(propensity) $P(C_1, C_2)$ 。

$$R_{con}(C_1, C_2) = \frac{f_{con}(C_1, C_2)/f_{all}(C_1, C_2)}{\sum_{i=1}^7 \sum_{j \neq i} f_{con}(C_i, C_j)/f_{all}(C_i, C_j)}$$

$$R_{non}(C_1, C_2) = \frac{f_{non}(C_1, C_2)/f_{all}(C_1, C_2)}{\sum_{i=1}^7 \sum_{j \neq i} f_{non}(C_i, C_j)/f_{all}(C_i, C_j)}$$

$$P(C_1, C_2) = \frac{R_{con}(C_1, C_2)}{R_{con}(C_1, C_2)}$$

其中 $f_{con}(C_1, C_2)$ 是相互作用的氨基酸对 $(x, y) \in C_1 \times C_2$  的频率。类似,  $f_{non}(C_1, C_2)$  和  $f_{all}(C_1, C_2)$  分别是不相互作用的氨基酸对的频率和所有可能的氨基酸对的频率。

- (3) 根据氨基酸对  $(C_1, C_2)$  附近氨基酸的分部情况, 计算权重  $w(C_i|(C_1, C_2))$ ,  $i = 1, 2, \dots, 7$

$$w_{con}(C_i|(C_1, C_2)) = \frac{f_{con}(C_i)/f_{all}(C_i)}{\sum_{C_i \in N} f_{con}(C_i)/f_{all}(C_i)} \chi_{\{C_i \in N\}}$$

$$w_{non}(C_i|(C_1, C_2)) = \frac{f_{non}(C_i)/f_{all}(C_i)}{\sum_{C_i \in N} f_{non}(C_i)/f_{all}(C_i)} \chi_{\{C_i \in N\}}$$

$$w(C_i|(C_1, C_2)) = w_{con}(C_i|(C_1, C_2))/w_{non}(C_i|(C_1, C_2))$$

其中  $f_{con}(C_i)$  为属于类别  $C_i$  的参与相互作用的氨基酸出现的频率,  $N$  为  $(C_1, C_2)$  附近氨基酸的类别的集合, 在这里, “附近” 可以根据具体情况选择的距离范围来决定。 $\chi$  为示性函数(characteristic function)。

- (4) 计算一个抗原表面的氨基酸  $x$  可以成为抗原表位的 i-Patch 分数。

$$\text{i-Patch}(x) = \frac{1}{|para|} \sum_{C_i \in N(C_x, C_y)} \sum_{y \in para} w(C_i|(C_x, C_y)) P(C_x, C_y)$$

其中  $para$  是给定抗体的抗体表位氨基酸的集合,  $C_x$  和  $C_y$  分别是  $x$  和  $y$  所在的氨基酸类,  $N(C_x, C_y)$  为氨基酸对  $(x, y)$  附近氨基酸类别的集合。“附近”的界定和 (3) 同。

从上面的计算可以看出, i-Patch 考虑了给定抗体的抗体表位, 同时也以权重的形式, 考虑了氨基酸对周围氨基酸的影响。

Liang Zhao 和 Jinyan Li 设计了 Bepar(B-cell epitope prediction through association rules)[42]。其大概方法如图1.7:

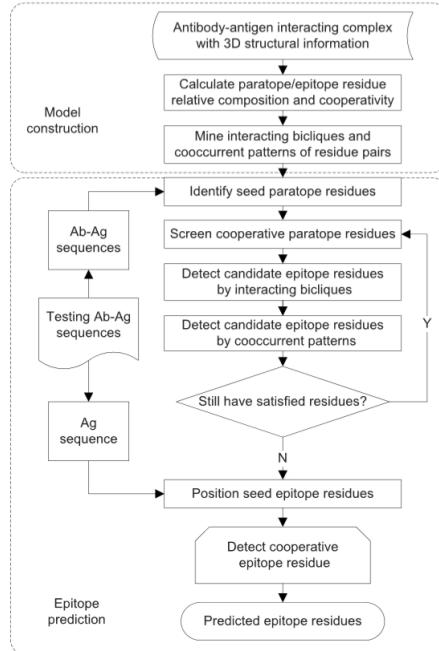


图 1.7: 此图源自文献[42]

其中涉及到抗原表位和抗体表位相互作用的部分有相对组成(relative composition)、Cooperativity 和 Cooccurrent pattern。它们的计算如下：

### (1) Relative composition 的计算

$$R_{ij} = 2 \times P_{ij} \times \log \frac{P_{ij}}{Q_{ij}}$$

其中， $ij$  表示第  $i$  个CDR上的氨基酸  $j$ 。 $P_{ij}$  表示氨基酸  $j$  在第  $i$  个CDR上所占的比例， $Q_{ij}$  表示氨基酸  $j$  在所有CDR上所占的比例。同样抗原表位氨基酸 Relative composition 的计算如下：

$$R_j = 2 \times P_j \times \log \frac{P_j}{Q_j}$$

其中  $P_j$  是抗原表位中氨基酸  $j$  在所有抗原表位氨基酸中的比例， $Q_j$  是氨基酸  $j$  在所有抗原氨基酸中的比例。

### (2) Cooperativity 的计算

$$C_{i,jk} = \frac{P_{i,jk}}{Q_{i,jk}}$$

$C_{i,jk}$  表示在第  $i$  个CDR上，两个相邻的氨基酸  $jk$  的 Cooperativity。 $P_{i,jk}$  表示两个相邻的氨基酸在  $jk$  在第  $i$  个CDR上所占的比例， $Q_{i,jk}$  表示两个相邻的氨基酸在  $jk$  在所有CDR上所占的比例。

- (3) Association Rule 的计算，具体可以参看文献[43]。大概的做法是把抗原表位和抗体表位的相互作用看成二部图然后寻找同一个抗原抗体复合物中出现的二部图的相似性。

Relative composition 被用来寻找种子，然后通过 Cooperativity 和 Association Rule 进行扩展。

Inbal Sela-Culang 等开发了基于给定抗体序列的抗原表位的预测方法：PEASE[]。该方法通过抽取氨基酸对的10个特征，然后构建 Random Forest 来预测。这10个特征分别是：

- (1) 氨基酸对中，来自抗原的氨基酸C端一侧相邻的4个氨基酸和N端一侧相邻的4个氨基酸。
- (2) 氨基酸对中，来自抗体的氨基酸C端一侧相邻的4个氨基酸和N端一侧相邻的4个氨基酸。
- (3) 氨基酸对中，来自抗原的氨基酸的 solvent accessibility。
- (4) 氨基酸对中，来自抗原的氨基酸的二级结构(secondary structure)。
- (5) 氨基酸对中，来自抗体的氨基酸是来自于轻链还是重链。
- (6) CDR的序号，是1、2还是3。
- (7) 氨基酸对中，来自抗体的氨基酸在CDR中的位置。
- (8) 氨基酸对中，来自抗原的氨基酸的无序状态(disorder state)。
- (9) 氨基酸对中，来自抗原的氨基酸的作用热点的分类(classification of the interaction hotspot)。
- (10) 氨基酸对的 Contact potential, 计算如下：

$$E_{ij} = \log_2 \left[ \frac{f^{ij}}{\sum_{i=1}^{20} \sum_{j=1}^{20} f^{ij}} \right] / \left( \frac{\sum_{j=1}^{20} f^j}{\sum_{i=1}^{20} f^i} \times \frac{\sum_{i=1}^{20} f^i}{\sum_{j=1}^{20} f^j} \right)$$

其中， $i$ 表示抗体表位上的氨基酸， $j$ 表示抗原表位上的氨基酸， $f^i$ 表示氨基酸*i*在抗体表位上出现的频率， $f^j$ 表示氨基酸*j*在抗原表位上出现的频率, $f^{ij}$ 表示氨基酸*i* 和 *j* 相互作用的频率， $E_{ij}$  为要计算的Contact potential 。

以上的这四种方法，是我能够了解到的基于抗体的抗原表位预测方法，虽然这些方法是为了预测抗原的表位，但是也可以看成对抗原抗体相互作用的探索。如果仔细观察这些方法，它们主要是对氨基酸对进行分析，也就是对1个抗原表位氨基酸和1个抗体表位氨基酸进行分析，而且这些分析多是基于出现的频率大小。并没有综合运用氨基酸与氨基酸之间的替换矩阵(substitution matrix)，也没有涉及到多个氨基酸与多个氨基酸之间的对应关系。鉴于此，本论文就是在这些方面做了改进，建立了一个更有效的模型。

蛋白质和蛋白质的相互作用主要集中在几个关键的氨基酸上，这几个关键的氨基酸被称为热点(hot-spot)[44]。最早提出热点的概念，是 Clackson 等在研究人生长激素(hGH)和它的一个受体hGHbp 相互作用时提出的[45]。通常来讲，这些热点上的每个氨基酸对亲和力的贡献都在 2kJ/mol 以上，而且相互作用的两个蛋白的热点是相互对应的，不但结构互补，亲水性也相同，带电荷也往往相反。系统的分析表明，热点的氨基酸组成个蛋白中氨基酸的组成有明显的不同[?]。Tryptophan(21%)，Arginine(13.3%) 和 Tyrosine(12.3%) 是含量最高的三种氨基酸。

抗原与抗体的相互作用作为蛋白质之间相互作用的一个特例，热点的概念也同样适用。多数的情况下，抗原抗体的亲和力是由少数几个热点氨基酸决定，比如文献[47]的研究。通过实验手段确定抗原抗体的热点，往往通过 Alanine Scanning。也有通过计算的方法来确定热点[48, 49]。

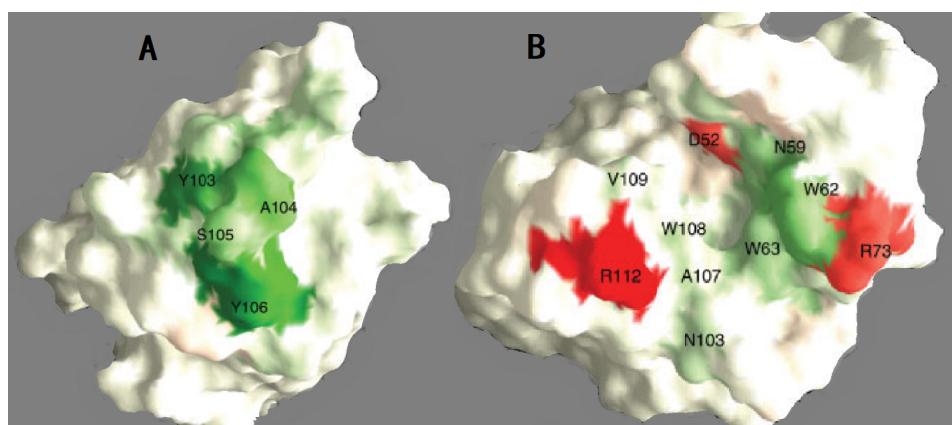


图 1.8: A 为抗体 cAB-Lys3 和上面的热点； B 为抗原 HEL(蛋清溶菌酶)上的热点。图片来自[49]，绿色代表加强抗原抗体的相互作用，红色代表减弱抗原抗体的相互作用。

从图1.8可以看出，热点只有少数的几个氨基酸，而且往往是短的在序列上相邻的氨基酸组成。在图1.8A 中热点为 Y103, A104, S105, Y106 。B 中也有类似情况。既然这样，如果我们只研究抗原热点和抗体热点之间的关系，就可以把问题在一定程度上化简为一个短的氨基酸序列和另外一个短的氨基酸序列的相互作用。

### 第3节 亲和力预测

自然产生的抗体，往往在亲和力(affinity)上有一定的限制[50]。为了获得更好的疗效，在治疗性抗体的生产和设计过程中，往往要涉及到亲和力成熟(affinity maturation)。所谓的亲和力的成熟就是说，在原来抗体的基础上，通过改造CDR上氨基酸的序列，从而使得抗原和抗体之间有更高的亲和力[51]。传统的实验手段，比如说突变和各种各样的展示技术，可以被用来寻找具有更高亲和力的抗体[52]，有不少的亲和力(affinity)可以被提高到nM的级别[53]。但是这些实验的手段，需要很大的工作量，也会持续很长的时间。也有一些通过计算的手段，来指导CDR上氨基酸的序列的改造。这样的方法也比较多，比如通过分子模拟、通过自由能的扰动(free energy perturbation)[54],通过 potential-of-mean force(PMF)[55]，虽然目前现在这些通过计算提高亲和力的方法效果还不是太理想，但是，这是一片广阔的天地[56]。首先，计算的方法可以大大提高速度。传统方法几个月才能完成的事情，计算的方法可能几天就可以完成。其次，随着数据量的积累和计算水平的提高，计算的结果必然会越来越准确。通过计算的方法解决亲和力的问题，概括的讲来，大概有两个思路，一个是从统计学的角度，一个是从物理化学原理的角度。上面说的分子模拟等方法都是从物理化学的角度，来进行的计算。也可以从统计学的角度来解决这个问题，比如说观察到观察到A和B经常相互作用，而A和C不经常相互作用。如果把C突变成B就很有可能增加分子的亲和力。Lippow等[57]，集中分析氨基酸与氨基酸之间的静电作用，高效的把抗EGFR的治疗性抗体Erbitux 的亲和力提高10倍，达到52pM；把抗蛋清溶菌酶的抗体D44.1 的亲和力提高了140倍，达到30pM；同时他们方法的有效性还在治疗性抗体Avastin 的已知亲和力的突变，以及抗荧光蛋白(fluorescein) 的抗体 4-4-20 的已知亲和力的突变中，得到验证。这个结果是惊人的，是计算机辅助亲和力成熟的一次成功。但是，这个方法需要抗原和抗体复合物的结构，同时也只分析了一对一的静电相互作用，并没有涉及到多个氨基酸协同作用和其他种类的相互作用力的情况。Sarah Sirin 等[58]构建了一个抗原抗体复合物突变数据库，涉及到32个复合物，1101个突变，并且每个突变都有实验手段得到的亲和力。通过比较 dDFIRE, DFIRE, STATIUM, Rosetta, FoldX, Discovery Studio, 在预测亲和力上的表现，Sarah Sirin 等认为，总的来说，目前的预测方法效果都不够好，实际  $\Delta\Delta G$  (吉布斯自由能的变化)和预测的  $\Delta\Delta G$  之间的相关系数  $r$  在 0.16 到 0.45 之间，很难满足实际的需求，但是在作出亲和力是增加还是减小这样的二分类判断时候，这些方法具有一定的意义。在上面的方法中，FoldX 和 Discovery Studio 是效果较好的两个。在上面的所有方法中，STATIUM 使用了统计学的方法，通过计算两个氨基酸之间  $C^\alpha$  和  $C^\beta$  之间的距离和夹角的关系，来计算氨基酸之间的 STATIUM potential，进而估计突变对亲和力的影响[59]。

## 第4节 本论文的主要内容

本论文可以概括为通过对热点的研究，来建立抗原抗体相互作用的模型。中间涉及的内容，从数据的收集，到模型的检验，还有不同模型的比较。可以成如下几点：

- (1) **数据的提取和初步分析。**在第一章最后的部分说，多数情况下，抗原抗体的相互作用由几个关键的氨基酸决定，这些关键的氨基酸往往是一些短的相邻的氨基酸。那么如何找出这些氨基酸？这是本论文解决的第一个问题。
- (2) **RBFN(径向基函数网络) 模型的构建。**找到了相互作用的短的氨基酸序列的相互作用关系后，如何刻画这个关系？本论文首先通过氨基酸替换矩阵(BLOSUM62)，定义了短的氨基酸序列对之间的距离，用径向基函数网络对这个关系进行了刻画。然后，对这个模型进行了检验。
- (3) **RBFN 模型的应用。**建好的模型可以用来做什么？本论文用这个模型做了两件事，一是通过模型来验证抗原表位和抗体表位的不同，二是通过模型来预测突变后，抗原抗体亲和力的变化。
- (4) **RBFN center 的选取。**在构建 RBFN 模型时候，一个重要的步骤就是选取合适的中心(center)。本论文设计了一个更为高效的选取方法，显著好于目前已知的方法。

## 第二章 数据的提取和初步分析

### 第1节 数据提取



图 2.1: 数据提取流程

- (1) **抗原抗体复合物的选取。** 抗原抗体复合物通过由 Oxford Protein Informatics Group 维护的网站下载[60]。下载时候设定复合物的分辨率不小于3Å 并且抗原的不小于5个氨基酸的长度。一共收集了1624个抗原抗体复合物，并把最近提交的 10% 的复合物作为 testing set。
- (2) **提取所有相互作用的氨基酸对。** 这里所有相互作用的氨基酸对，是指在CDR上的所有的氨基酸和抗原上的氨基酸之间所有可能的相互作用。这里涉及到CDR区域的界定和相互作用的定义。
  - (a) CDR 区域的定义方法比较多[61]，它们有一定程度上的差别，但是总体一致。 Andrew C. R. Martin 教授的团队在给出了确定CDR区域不同方法之间的差别，并给出了一些判断准则([www.bioinf.org.uk/abs](http://www.bioinf.org.uk/abs))。单单从序列上看，轻链上面的CDR 区域的判断规则为：
    - i. 第一个CDR(CDRL1) 大概从第24位开始，长为10-17。
    - ii. 第二个CDR(CDRL2) 从CDR1后的第16个氨基酸始，长为7。
    - iii. 第三个CDR(CDRL3)从CDR2后的第33个氨基酸始，长为7-11。为了涵盖可能的CDR区域， CDRL1的位置，其区域应该设定为24-41，同样CDRL2的区域设定为50-64， CDRL3的区域设定为90-108。可以总结为表格2.1： 上面的定义方式，在包含轻链上所有可能CDR 的同时，也会包含非CDR区域。但是这个对后面数据的分析很小，一则因为抗原抗体的相互作用多数发生在CDR区域和抗体之间，二则当后面抽取最关键的氨基酸时，又进一步的把范围限制在CDR上，所以非CDR区域的氨基酸，即便被包含

	CDRL1	CDRL2	CDRL3
Starting Length	Approximately 24 10 to 17 residues	16 residues after CDRL1 7 residues	33 residues after CDRL2 7 to 11 residues
Max-CDRL	24 to 41	50 to 64	90 to 108

表 2.1: 轻链上CDR区域的确定

进去，后面被选来作为数据的可能性也很小。所以，这里宁愿把CDR的范围定大，也不定小。同样的原因，我们可以确定重链上CDR的范围(表2.2)。

	CDRH1	CDRH2	CDRH3
Starting Length	Approximately 26 10 to 12 residues	15 residues after CDRH1 16 to 19 residues	33 residues after CDRH2 3 to 25 residues
Max-CDRH	26 to 38	51 to 72	100 to 130

表 2.2: 轻链上CDR区域的确定

(b) **相互作用氨基酸对的确定。**如果分别位于CDR 和抗原上的两个氨基酸，存在距离不大于  $4\text{\AA}$  的原子(氢原子除外)，那么就认为这两个氨基酸相互作用。为了便于这里的陈述和后面进一步的应用，我们定义两个氨基酸间的作用数(contact number)CN:

$$\text{CN}(A, B) = \sum_{a \in A} \sum_{b \in B} \chi\{d(a, b) \leq 4\} \quad (2.1)$$

其中A，B 分别表是两个氨基酸，在  $a \in A$  中，A 表示氨基酸 A 上除氢原子外所有原子的集合； $a \in A$  做同样解释。 $d(a, b)$  表示原子 a 和 b 之间的欧式距离，以  $\text{\AA}$  为单位。CDR 上的氨基酸 A 和抗原上的氨基酸 B 是相互作用的氨基酸对，则是说  $\text{CN}(A, B) > 1$ 。

确定了CDR的范围并定义了相互作用的氨基酸对之后，提取所有相互作用的氨基酸对则可以顺利进行。看一个例子。

(3) **去除重复。**为了避免同一个抗原抗体复合物被反复抽取数据，需要对复合物去重。然而，同一个抗原可以被不同的抗体识别，所以不能从抗原的角度进行性去重。从抗体的结构上可以知道，抗体和抗体之间的差别主要体现在可变区上，更进一步的说，主要体现在CDR区域上，所以，只要两个抗体的CDR区域差别足够大，就可以认定为不同的抗体，也就是，要从CDR区域的相似度上去重。这里，

我们把同一个抗体上的轻链和重链分别作为独立的两条链来考虑。设 LA, LB 分别为两条轻链。首先把 LA, LB 链上的三个CDR区域分别拼接起来，分别记为 CDR-LA, CDR-LB，然后进行比较，把打分规则定义为：

- (a) 相同的两个氨基酸得分为1
- (b) 不同的两个氨基酸得分为0
- (c) 空缺(gap)的得分为0
- (d) extended gap 得分为0

上面的规则可以不严格的概括为，计算相同氨基酸的数目。把通过上面的规则得到的分数记为  $S(LA, LB)$ 。然后计算 LA 和 LB 之间的距离  $D(LA, LB)$ 。

$$D(LA, LB) = 1 - S(LA, LB)/N$$

其中  $N = \min(\text{Len}(\text{CDR-LA}), \text{Len}(\text{CDR-LB}))$ , Len 表示计算氨基酸序列的长度。上面的公式可以简单的该概括为，对  $S(LA, LB)$  进行 Normalization，然后转化为距离。

定义好距离，我们就根据这个距离对不同的轻链进行聚类(hierarchical cluster)，通过分析在不同截断距离 (cut-off distance) 下的聚类情况来决定去重的截断距离。图2.2是分类数目和截断距离之间的关系。

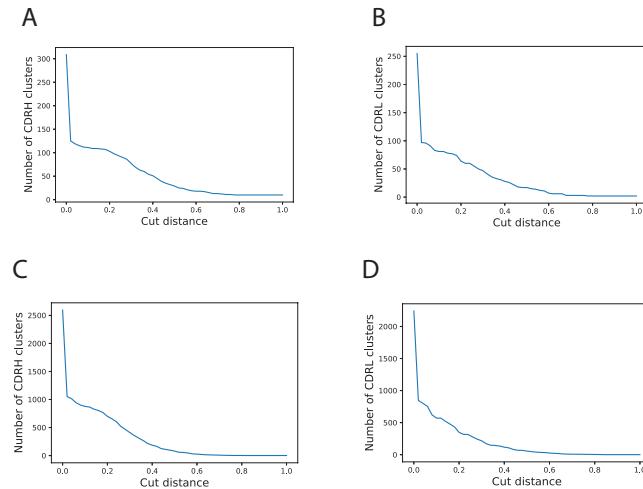


图 2.2: 截断距离和聚类数之间的关系。A 为 testing set 中重链的聚类情况，B 为 testing set 中轻链的聚类情况，C 为 training set 中重链的聚类情况，D 为 training set 中轻链的聚类情况。

从图中可以看出，当截断距离是0.1时，聚类数目的减少速度明显减缓。根据 elbow method, 0.1是一个比较好的截断距离。

当重链和轻链按照 0.1 的截断距离进行分类后，需要从每个类别中选出一个代表，作为有效数据使用。选取规则定义如下：

$$T(C) = \sum_{A \in C} \sum_{B \in Ag} CN(A, B)$$

其中  $C$  为一个轻链或者重链，在  $A \in C$  中  $C$  表示轻链或者重链CDR上氨基酸的集合，  $CN$  的定义见公式(2.1)。在每个分类中选取代表性的重链或者轻链的时候，选择  $T(C)$  最大的一个。表格2.3 给出了被选中的一个重链和其对应的抗原之间的所有相互作用的氨基酸对。

标记	抗体氨基酸位置	抗原氨基酸位置	CN
h1HA	30	16	7
h1HA	30	14	3
h1HA	30	17	7
h2HA	52	16	9
h2HA	52	15	2
h3HA	99	196	1
h3HA	99	198	4
h3HA	100	196	3
h3HA	101	196	8
h3HA	101	15	4
h3HA	101	13	1
h3HA	101	14	3
h3HA	101	190	3
h3HA	101	198	6
h3HA	102	196	5
h3HA	102	197	3

表 2.3: 来自于抗原抗体复合物 1adq(一个PDB编号) 重链和其对应抗原之间所有坑能相互作用的氨基酸。上面使用的是本论文涉及的四坐标体系(four coordinate system)，这四个坐标分别是 标记，抗体氨基酸位置，抗原氨基酸位置 和 CN。标记，给出抗体链的名称，抗原链的名称和CDR序号；CN如公式(2.1)定义。(h1HA, 30, 16, 7)表示在复合物 1adq 的重链 H 上的 30 位氨基酸和抗原链 A 上 16 位氨基酸相互作用，请 CN 值为 7，并且这个相互作用位于重链的第一个CDR上。

- (4) **组合成相互作用的氨基酸序列** 既然最终的目的是要选出其关键作用的短的相互作用的氨基酸序列，那么就要把表2.3 中所有可能相互作用的氨基酸按照不同的长度连连接起来。先来定义一个概念：匹配类型(match-type)。所谓匹配类型就是相互作用的连续的抗体氨基酸的长度和相互作用的连续的抗原氨基酸的长度所组成的一个二维坐标。比如 (2,3) 就表示 2 个连续的抗体氨基酸和 3 个连续的抗原氨基酸相互作用。在本论文中匹配类型的取值范围是 $\{(i, j) : i, j = 1, 2, 3\}$ 。对表2.3中的数据进行组合：
- (a) 匹配类型为(1,1)的有: ([30],[16],7), ([30], [14],3), ([30], [17], 7), ([52], [16], 9), ([52], [15], 2), ([99], [196], 1), ([99], [198], 4), ([100], [196], 3), ([101], [196], 8), ([101], [15], 4), ([101], [13], 1), ([101], [14], 3), ([101], [190], 3), ([101], [198], 6), ([102], [196], 5), ([102], [197], 3)。这里最后一个坐标为CN值，方便下一步Core的选取。([30],[16],7) 表示抗体上30位的氨基酸和抗原上16位的氨基酸相互作用，其CN值为7。
  - (b) 匹配类型为(1,2)的有: ([30], [16,17], 14), ([52], [15, 16], 11), ([101], [13, 14], 4), ([101], [14, 15], 7), ([102], [196, 197], 8)。
  - (c) 匹配类型为(1,3)
  - (d) 匹配类型为(2,1)
  - (e) 匹配类型为(2,2)
  - (f) 匹配类型为(2,3)
  - (g) 匹配类型为(3,1)
  - (h) 匹配类型为(3,2)
  - (i) 匹配类型为(3,3)
- (5) **选出 core** 通过计算的方法找关键的相互作用的氨基酸方法比较多，有的从物理化学的角度进行计算，有的计算接触氨基酸之间接触面积的大小。本论文通过计算CN的大小来判断一个氨基酸是不是起着关键作用。

## 第 2 节 数据分析

使用三号字，黑体，居中对齐。

## 第 3 节 节标题

使用小三号字，黑体，居中对齐。

## 第 4 节 子节标题

使用小四号字, 黑体, 靠左对齐.

## 第 5 节 正文

使用小四号字, 行距为20磅. 首行缩进两个字符宽. 建议标点符号用半角. 例如句号用“句点”. 输入时每个标点后打一个空格.

## 第 6 节 章节

如果文章内容较多, 可以采用分章节. 如果内容较少, 可以只用节而不用章. 章节的编号方式(编号类型等的选择)要恰当.

## 第三章 公式排版

这部分介绍如何正确使用公式编排.

$$F(b) - F(a) = \int_a^b F'(x) dx. \quad (3.1)$$

### 第 1 节 行内公式

如果  $x = y, y = z$ , 那么我们可以推得  $x = z$ . 如果式子过长, 应该写成行间公式.

### 第 2 节 行间公式

如果  $x = y$ , 那么

$$f(x) = f(y)$$

但是, 若  $x \neq y$ , 我们也不能获得

$$f(x) \neq f(y) \quad (3.2)$$

所以 (3.2) 不是  $x \neq y$  的必要条件.

下面是另外的例子: 第一个公式不标号, 请注意命令\nonumber的使用:

$$\begin{aligned} W_{i,a}^{\text{new}} &\leftarrow W_{i,a} \sum_{\mu} \frac{V_{i,\mu}}{(WH)_{i,\mu}} H_{a,\mu} \\ H_{a,\mu}^{\text{new}} &\leftarrow H_{a,\mu} \sum_i W_{i,a} \frac{V_{i,\mu}}{(WH)_{i,\mu}} \end{aligned} \quad (3.3)$$

$$W_{i,a}^{\text{new}} \leftarrow \frac{W_{i,a}}{\sum_j W_{j,a}} \quad (3.4)$$

如果所有公式都不标号, 可以采用下面的环境:

$$\begin{aligned}
 (\arcsin x)^2 &= \left( \sum_{k=0}^{\infty} \frac{C_{2k}^k}{2k+1} \frac{x^{2k+1}}{2^{2k}} \right)^2 \\
 &= \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} \frac{C_{2k}^k C_{2j}^j}{(2k+1)(2j+1)} \frac{x^{2k+2j+2}}{2^{2k+2j}} \\
 &= \sum_{n=0}^{\infty} \sum_{k+j=n} \frac{C_{2k}^k C_{2j}^j}{(2k+1)(2j+1)} \frac{x^{2n+2}}{2^{2n}} \\
 &= \sum_{n=0}^{\infty} \frac{(2x)^{2n+2}}{2C_{2n+2}^{n+1}(n+1)^2}.
 \end{aligned}$$

更多公式环境的使用以及一些数学符号的使用可以参考一些`LATEX`的书籍.

本模板中, 在每章开头, 公式标号重新计数. 一章中, 即使换节, 计数并不重新开始(比较(3.1), (3.2)), 请注意公式编号的引用以及对应的超链接效果.

若各节的公式需要重新编号, 可自行修改, 比如利用命令

```
\def\theequation{\arabic{chapter}.\arabic{section}.\arabic{equation}}
```

(或 `\def\theequation{3.2.\arabic{equation}}`)

```
\setcounter{equation}{0}
```

利用以上命令也可以解决诸如引入带撇的编号“3.1.3'”, 以及回到正常编号的重新编号问题.

上述命令下的公式编号:

$$\lim_{n \rightarrow +\infty} \left(1 + \frac{1}{n}\right)^n = e. \quad (3.2.1)$$

定义、定理、例子等的编号格式也可以用类似命令.

## 第四章 表格和图片

Dataset	Before	After	Percentage
ALL/AML leukaemia	7129	1038	14.56
Breast Cancer	24 481	834	3.41
CNS embryonal tumous	7129	74	1.04
Colon tumour	7129	135	1.89
Lung cancer	12 533	5365	42.81
Prostate cancer	12 600	3071	24.37
outcome	12 600	208	1.65

表 4.1: 这是个表格

如果插图, 可以考虑下面的命令:

```
\includegraphics[options]{yourfile}
```

具体命令参考 `graphicx` 宏包说明, 值得注意的是用 PDF LATEX 编译是不支持插入 EPS 格式图片的, 不过将 EPS 格式图片转换为 PDF 后就可以插入了. 限于条件限制, 本模板不给出插入图片的示例.

论文中的数据图例可以由 MatLab 制作 (比如数据模拟图), 一般的图例 (含流程图, 交换图等) 可由 MetaPost 或者 Asymptote 作出 (当然作图工具不限于此), 限于条件限制, 模板不给出示例.

## 第五章 定理环境

### 第 1 节 题头

同一章内定理、引理等“题头”可以采用连续/统一的标号，这是由模板中的诸如“`\newtheorem{theorem}[definition]{定理}`”这样的命令中的“`[definition]`”选项确定的，它使所有定理采用和定义统一编号：

引理 5.1. 对于任何实数  $A$ , 成立着  $A^2 \geq 0$ .

定理 5.2. 设  $A, B$  是两个实数, 则  $2AB \leq A^2 + B^2$ .

### 第 2 节 同章另一节的题头

推论 5.3. 设  $a, b$  为两个正数, 则其几何平均不大于其算术平均, 即  $\sqrt{ab} \leq \frac{a+b}{2}$ .

## 第六章 参考文献的写法

所有参考文献均用尾注形式列在论文篇末, 内容包括: 主要负责人(作者, 编者) 文献题名. 出版地, 出版年份, 起止页码. (如果文献是期刊杂志内的文章, 则除要列出作者和题名外, 还要注明期刊名, 出版时间, 卷号或期号, 起止页码).

英文出版物见[62], 国际会议见[64], 英文期刊见[63].

中文出版物见[65], 中文期刊见[66].

建议文献排序按作者姓氏的字母排序, 同一作者的文章按时间先后排列. 英文姓名的写法有先姓后名([67])和先名后姓([63])两种写法, 请统一到其中一种.

注意“参考文献”不写成论文的一章.

## 致谢

请对帮助过你完成论文的老师、同学致谢。也可以在此对您四年大学生活有重要帮助的人致谢。

“致谢”本身不作为一章，致谢内容的字体大小不宜与作为标题的“致谢”两字的大小有很大的反差。这一点尤其请使用word模板的同学注意。一般说来，杂志论文的致谢在文章正文结束、参考文献前(即本模板中它所处的位置)；学位论文的致谢在最后一页，并宜单独成页；书籍的致谢在序言结尾。

## 参考文献

- [1] Gary J. Nabel, Designing Tomorrow's Vaccines, *N Engl J Med*, 6(2013), 368:551-560.
- [2] Peter D. Kwong, John R. Mascola and Gary J. Nabel, Broadly neutralizing antibodies and the search for an HIV-1 vaccine: the end of the beginning, *Nat Rev Immunol*, (2013), 13:693-701.
- [3] <https://www.antibodysociety.org/resources/approved-antibodies/>
- [4] Drew M. Pardoll, The blockade of immune checkpoints in cancer immunotherapy, *Nat Rev Cancer*, (2012), 12(4): 252 - 264.
- [5] António L.Grilo, A.Mantalaris,The Increasingly Human and Profitable Monoclonal Antibody Market, *Trends in biotechnology* 37.1 (2019): 9-16.
- [6] Padlan, Eduardo A., et al. Structure of an antibody-antigen complex: crystal structure of the HyHEL-10 Fab-lysozyme complex. *PNAS*, (1989), 86(15):5938-5942.
- [7] Yokota, Akiko, et al. The Role of Hydrogen Bonding via Interfacial Water Molecules in Antigen-Antibody Complexation THE HyHEL-10-HEL INTERACTION. *Journal of Biological Chemistry*, (2003), 7(278):5410-5418.
- [8] Yokota, Akiko, et al. Contribution of asparagine residues to the stabilization of a proteinaceous antigen-antibody complex, HyHEL-10-hen egg white lysozyme. *Journal of Biological Chemistry*, (2010), 285(10): 7686-7696.
- [9] Pons Jaume, Arvind Rajpal, and Jack F. Kirsch. Energetic analysis of an antigen/antibody interface: alanine scanning mutagenesis and double mutant cycles on the HyHEL-10/lysozyme interaction. *Protein Science*,(1999), 8(5): 958-968.
- [10] Shiroishi, Mitsunori, et al. Structural Consequences of Mutations in Interfacial Tyr Residues of a Protein Antigen-Antibody Complex THE CASE OF HyHEL-10-HEL. *Journal of Biological Chemistry*,(2007), 282(9): 6783-6791.
- [11] Shiroishi, Mitsunori, et al. Structural Evidence for Entropic Contribution of Salt Bridge Formation to a Protein Antigen-Antibody Interaction THE CASE OF HEN LYSOZYME-HyHEL-10 Fv COMPLEX. *Journal of Biological Chemistry*, (2001), 276(25): 23042-23050.
- [12] Kam-Morgan, L. N., et al. High-resolution mapping of the HyHEL-10 epitope of chicken lysozyme by site-directed mutagenesis. *PNAS*,(1993),90(9): 3958-3962.
- [13] Novotny, Jiri, Robert E. Bruccoleri, and Frederick A. Saul. On the attribution of binding energy in antigen-antibody complexes McPC 603, D1. 3, and HyHEL-5. *Biochemistry*,(1989), 28(11): 4735-4749.
- [14] Li, Yili, et al. Three-dimensional structures of the free and antigen-bound Fab from monoclonal antilysozyme antibody HyHEL-63. *Biochemistry*, (2000), 39(21): 6296-6309.
- [15] Hibbits, Kari A., Davinder S. Gill, and Richard C. Willson. Isothermal titration calorimetric study of the association of hen egg lysozyme and the anti-lysozyme antibody HyHEL-5. *Biochemistry*,(1994),33(12): 3584-3590.

- [16] Cohen, GERsoN H., S. Sheriff, and DAVID R. Davies. Refined structure of the monoclonal antibody HyHEL-5 with its antigen hen egg-white lysozyme. *Acta Crystallographica Section D: Biological Crystallography*, (1996), 52(2): 315-326.
- [17] Li, Yili, et al. Dissection of binding interactions in the complex between the anti-lysozyme antibody HyHEL-63 and its antigen. *Biochemistry*, (2003), 42(1): 11-22.
- [18] Xavier, K. Asish, et al. Involvement of water molecules in the association of monoclonal antibody HyHEL-5 with bobwhite quail lysozyme. *Biophysical journal*, (1997), 73(4) : 2116-2125.
- [19] Slagle, S. P., R. E. Kozack, and S. Subramaniam. Role of electrostatics in antibody-antigen association: anti-hen egg lysozyme/lysozyme complex (HyHEL-5/HEL). *Journal of Biomolecular Structure and Dynamics*, (1994), 12(2) : 439-456.
- [20] Cohen, Gerson H., et al. Water molecules in the antibody - antigen interface of the structure of the Fab HyHEL-5 - lysozyme complex at 1.7 Å resolution: comparison with results from isothermal titration calorimetry. *Acta Crystallographica Section D: Biological Crystallography*, (2005), 61(5): 628-633.
- [21] Wibbenmeyer, Jamie A., et al. Salt links dominate affinity of antibody HyHEL-5 for lysozyme through enthalpic contributions. *Journal of Biological Chemistry*, (1999), 274(38) : 26838-26842.
- [22] Ekiert, Damian C., et al. Antibody recognition of a highly conserved influenza virus epitope. *Science*, (2009), 324(5924): 246-251.
- [23] , S Ferdous, S Kelm, TS Baker, J Shi, ACR Martin, B-cell epitopes: Discontinuity and conformational analysis *Molecular immunology*, (2019), 114:643-650.
- [24] G. E. Morris, Epitope mapping, *Methods in Molecular Biology*, (2005), 295:255 – 268.
- [25] Rubinstein, Nimrod D., et al. Computational characterization of B-cell epitopes. *Molecular immunology*, (2008), 45(12):3477-3489.
- [26] Hopp, T. P., Woods, K. R. Prediction of protein antigenic determinants from amino acid sequences. *PNAS*, (1981), 78(6), 3824-3828.
- [27] Rubinstein, N. D., Mayrose, I., Pupko, T. A machine-learning approach for predicting B-cell epitopes. *Molecular immunology*, (2009), 46(5), 840-847.
- [28] Sun, J., Wu, D., Xu, T., Wang, X., Xu, X., Tao, L., ..., Cao, Z. W., SEPPA: a computational server for spatial epitope prediction of protein antigens. *Nucleic acids research*, (2009), 37(suppl\_2), W612-W616.
- [29] Kulkarni-Kale, U., Bhosle, S., Kolaskar, A. S., CEP: a conformational epitope prediction server. *Nucleic acids research*, (2005), 33(suppl\_2), W168-W171.
- [30] Haste Andersen, P., Nielsen, M., Lund, O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Science*, (2006), 15(11), 2558-2567.
- [31] Moreau, V., Fleury, C., Piquer, D., Nguyen, C., Novali, N., Villard, S., ..., Molina, F. PEPOP: computational design of immunogenic peptides. *Bmc Bioinformatics*, (2008), 9(1), 71.
- [32] Ponomarenko, J., Bui, H. H., Li, W., Fuseder, N., Bourne, P. E., Sette, A., Peters, B. ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC bioinformatics*, (2008), 9(1), 514.
- [33] 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.

- [34] Xu, X., Sun, J., Liu, Q., Wang, X., Xu, T., Zhu, R., ... Cao, Z., Evaluation of spatial epitope computational tools based on experimentally-confirmed dataset for protein antigens. *Chinese Science Bulletin*, (2010), 55(20), 2169-2174.
- [35] Sela-Culang, Inbal, Yanay Ofran, and Bjoern Peters. Antibody specific epitope prediction—emergence of a new paradigm. *Current opinion in virology*, (2015), 11 : 98-102.
- [36] Soga, S., Kuroda, D., Shirai, H., Kobori, M., Hirayama, N., Use of amino acid composition to predict epitope residues of individual antibodies. *Protein Engineering, Design & Selection*, (2010). 23(6), 441-448.
- [37] Krawczyk, K., Liu, X., Baker, T., Shi, J., Deane, C. M. Improving B-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics*, (2014), 30(16), 2288-2294.
- [38] Sela-Culang, I., Ashkenazi, S., Peters, B., Ofran, Y. PEASE: predicting B-cell epitopes utilizing antibody sequence. *Bioinformatics*, (2015), 31(8), 1313-1315.
- [39] Krawczyk, K. , Liu, X. , Baker, T. , Shi, J. , Deane, C. M.,Improving b-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics*, (2014), 30(16), 2288-2294.
- [40] Konrad Krawczyk1, Terry Baker, Jiye Shi, Charlotte M.Deane, Antibody i-Patch prediction of the antibody binding site improves rigid local antibody - antigen docking, *Protein Engineering, Design & Selection*, 26(2013), no.10, pp. 621–629.
- [41] Hamer, R. , Luo, Q. , Armitage, J. P. , Reinert, G. , Deane, C. M. , I-patch: interprotein contact prediction using local network information. *Proteins Structure Function & Bioinformatics*, (2010), 78(13), 2781-2797.
- [42] Zhao, L. , Li, J., Mining for the antibody-antigen interacting associations that predict the b cell epitopes. *BMC Structural Biology*, (2010), 10 Suppl 1(Suppl 1), S6.
- [43] Coenen F, Goulbourne G, Leng P, Tree Structures for Mining Association Rules. *Data Min. Knowl. Discov.*, (2004), 8:25-51
- [44] Moreira I S , Fernandes P A , Ramos M J . Hot spots—A review of the protein – protein interface determinant amino-acid residues. *Proteins*, 2007, 68(4):803-812.
- [45] Clackson T , Wells J . A hot spot of binding energy in a hormone-receptor interface. *Science*, (1995), 267(5196):383-386.
- [46] Keskin O, Ma B, Nussinov R. Hot regions in protein – protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol* 2005;345:1281 – 1294.
- [47] Dall Acqua W , Goldman E R , Lin W , et al. A Mutational Analysis of Binding Interactions in an Antigen-Antibody Protein-Protein complex, *Biochemistry*, (1998), 37(22):7981-7991.
- [48] Moreira, I.S., Fernandes, P.A. and Ramos, M.J., Hot spot computational identification: Application to the complex formed between the hen egg white lysozyme (HEL) and the antibody HyHEL-10. *Int. J. Quantum Chem.*, (2007), 107: 299-310.
- [49] Lafont, V., Schaefer, M., Stote, R.H., Altschuh, D. and Dejaegere, A., Protein-protein recognition and interaction hot spots in an antigen-antibody complex: Free energy decomposition identifies “efficient amino acids” . *Proteins*, (2007), 67: 418-434.
- [50] Foote, J., Eisen, H.N. Kinetic and affinity limits on antibodies produced during immune responses. *Proc. Natl. Acad. Sci*, (1995), 92, 1254-1256.
- [51] 5 Presta LG, Selection, design, and engineering of therapeutic antibodies. *J Allergy Clin Immunol*, (2005), 116(4):731-736

- [52] Chames P, Coulon S, Baty D. Improving the affinity and the fine specificity of an anti-cortisol antibody by parsimonious mutagenesis and phage display. *J Immunol*, (1998), 161(10):5421 - 9.
- [53] Lee CV, Liang WC, Dennis MS, Eigenbrot C, Sidhu SS, Fuh G. High-affinity human antibodies from phage-displayed synthetic Fab libraries with a single framework scaffold. *J Mol Biol*, (2004), 340(5):1073 - 93.
- [54] Chipot C, Pohorille A. Calculating Free Energy Differences Using Perturbation Theory. Free Energy Calculations *Springer Series in CHEMICAL PHYSICS: Springer*, (2007). p. 33 - 75.
- [55] Roux B. The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, (1995);91(1 - 3):275 - 82
- [56] Cannon DA, Shan L, Du Q, et al. Experimentally guided computational antibody affinity maturation with de novo docking, modelling and rational design. *PLoS Comput Biol*, (2019), 15(5):e1006980.
- [57] Lippow S M , Wittrup K D , Tidor B . Computational design of antibody-affinity improvement beyond in vivo maturation. *Nature Biotechnology*, (2007), 25(10):1171-1176.
- [58] AB-Bind: Antibody binding mutational database for computational affinity predictions. *Protein Science*, (2016), 25(2):393-409.
- [59] DeBartolo J, Dutta S, Reich L, Keating AE, Predictive Bcl-2 family binding models rooted in experiment or structure. *J Mol Biol* (2012), 422:124 - 144.
- [60] Dunbar,J., Krawczyk, K.,Leem, J.,Baker, T.,Fuchs, A.,Georges,G., Jiye Shi and Deane, C. M. SAbDab: the structural antibody database. *Nucleic acids research*, (2013), 42, D1140 - D1146.
- [61] Allazikani B, Lesk A M, Chothia C. Standard conformations for the canonical structures of immunoglobulins. *Journal of Molecular Biology*, (1997), 273(4):927-948.
- [62] T. Hastie et al., The Element of Statistical Learning, Springer Series in Statistics, Springer-Verlag, 2001.
- [63] S. Chen, Mach configuration in pseudo-stationary compressible flow, *J. Amer. Math. Soc.*, 21(2008), no. 1, pp. 63–100.
- [64] Junping Zhang, Li He, and Zhi-Hua Zhou, “Analyzing Magnification Factors and Principal Spread Directions in Manifold Learning”, in *Proceedings of the 9th Online World Conference on Soft Computing in Industrial Applications (WSC9)*, 2004.
- [65] 陈纪修, 潘崇华, 金路, 数学分析, 高等教育出版社, 1999.
- [66] 苏步青, 数学教育与应用数学问题, 数学通报, 1988, (2): 1–2.
- [67] Li, T. and Chen, Y., Global classical solutions for nonlinear evolution equations, Pitman Monographs and Surveys in Pure and Applied Mathematics, 45, Longman Scientific & Technical, Harlow.