

目录

第一章 研究的意义和现状	1
第 1 节 研究的意义	1
第 2 节 抗原抗体相互作用研究现状	5
第 3 节 本论文的主要内容	13
第二章 数据的提取和初步分析	15
第 1 节 数据提取	15
第 2 节 数据的初步分析	20
第三章 RBFN 的构建	26
第 1 节 Core 之间距离的定义	26
第 2 节 RBFN模型	27
第四章 RBFN 模型的应用	33
第 1 节 Core中抗原氨基酸和抗体氨基酸的区别	33
第 2 节 预测突变对抗原抗体复合物亲和力的影响	37
2.1. 可预测氨基酸对的提取	39
2.2. 亲和力变化的预测	40
第五章 讨论	44

抗原抗体相互作用的一个 RBFN 模型

刘传省

学号：17210180030

专业：计算系统生物学

摘要 抗原抗体的相互作用在人的免疫系统中扮演这至关重要的角色。对抗原抗体相互作用的研究，在治疗性抗体的设计、疫苗的生产上都有重要的意义。然而抗原抗体相互作用的空间性质，使得研究十分困难。本论文通过寻找抗原抗体相互作用的关键氨基酸序列，本论文称之为 Core 在一定程度上克服了这一困难。通过对 Core 中氨基酸的分析，发现 TYR 在抗体表位中的重要性，也为构建简化的抗体库提供了新的参考；同时发现抗体表位对亲水性氨基酸的偏好和疏水性氨基酸的排斥现象。本文在 Core 的基础上建立了 RBFN 模型来系统的描述抗原抗体之间相互作用的数量关系，取得很好效果。同时，这个模型被用来区分 Core 中的抗原氨基酸和抗体氨基酸，以及预测突变对抗原抗体复合物亲和力的影响，都取得了不错的效果。

关键字：抗原抗体相互作用，抗原表位，抗体表位，亲和力预测，Core，RBFN 模型，简化抗体库。

Abstract The interaction between antibody and antigen plays an important role in the immune system, and the study of those interactions benefit the design of therapeutic antibodies and the vaccine significantly. However the conformational property of the antibody-antigen interaction poses a big hurdle. To jump over this hurdle, we introduced the concept of the Cores, which were the key interacting amino acids, and by analyzing those Cores, the importance of TYR was emphasized, a new guide line for building antibody libraries with a few types of amino acids was suggested, and the favor of hydrophilic amino acids and repulsion of the hydrophobic amino acids of the epitopes were substantiated. An RBFN model was successfully built to systematically and numerically describe those cores. This model was exploited to distinguish between antibody amino acids and the antigen amino acids in the cores, and, more importantly, was exploited to predict whether a mutation could increase or decrease the affinity of an antibody-antigen complex.

Keywords: antibody-antigen interaction, paratope, epitope, affinity prediction, Core, RBFN model, reduced antibody library.

第一章 研究的意义和现状

第1节 研究的意义

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。

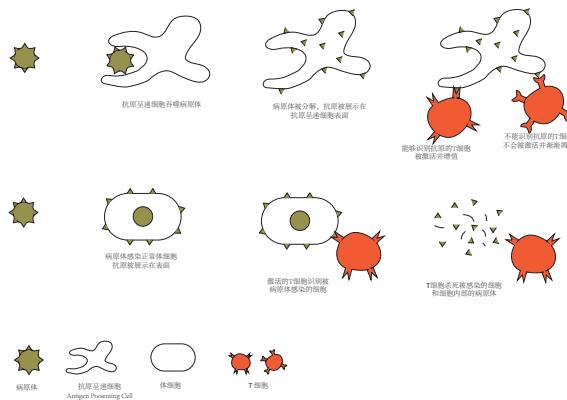


图 1.1: 特异性细胞免疫示意图。真实的过程比这要复杂的多，涉及的分子和细胞也比这多，但是这幅图可以说明基本过程。

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，称为抗原(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。

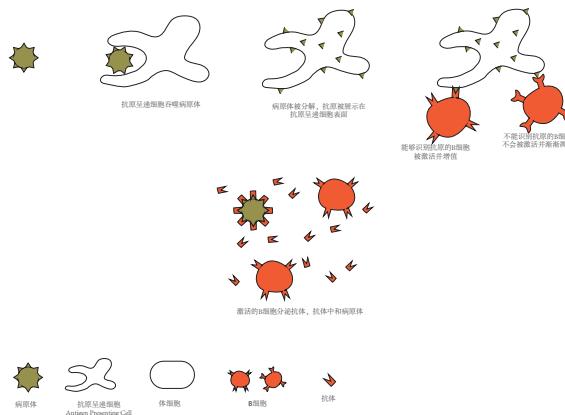


图 1.2: 特异性体液免疫示意图。此图简要说明了体液免疫过程中，抗体的产生和对病原的识别，真实的过程比这要复杂的多，比如说 T helper 的作用等。同时，抗体除了直接杀死病原体之外，还可以参与抗体介导的细胞毒性(antibody directed cell cytotoxicity)。

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。

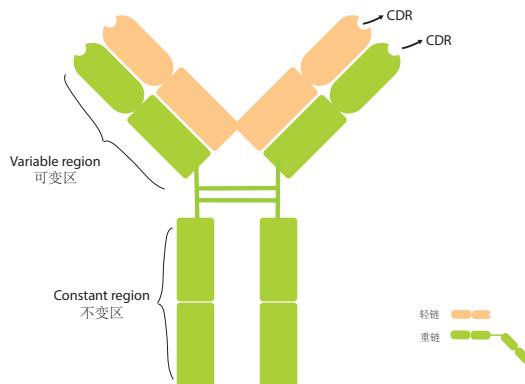


图 1.3: 抗体结构示意图

互补决定区主要由6个比较短的氨基酸片段组成，它们分为是来自重链和轻链的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 CDR 和抗原局部区域的相

互作用。抗原上那些和抗体相互作用的部分又称为抗原表位(epitope)。抗体上和抗原相互作用的部分称为抗体表位(paratope)。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。然而接种人痘，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。

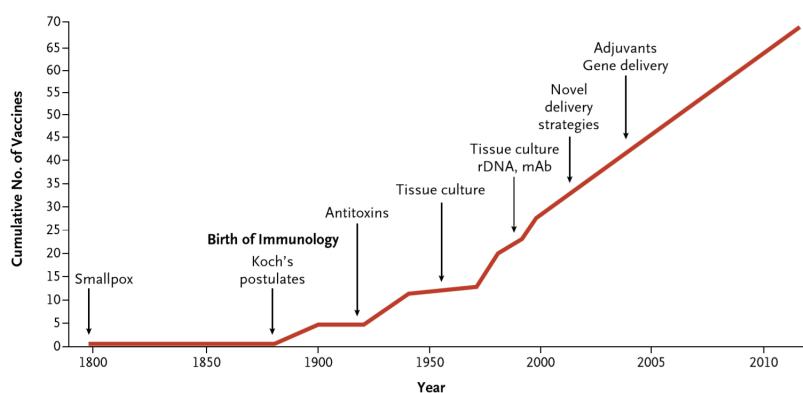


图 1.4: 疫苗的数量变化和疫苗开发技术的发展[1]

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 上一些保守的的抗原表位(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。要嫁接抗原表位，首先就要找到抗原表位。抗原表位的确定除了可以通过实验手段的到，还可以通过对抗原抗体相互作用规律的研究，由计算得到。

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]。目前，已经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，以及抗体的多样性和相关技术的发展，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1.5：

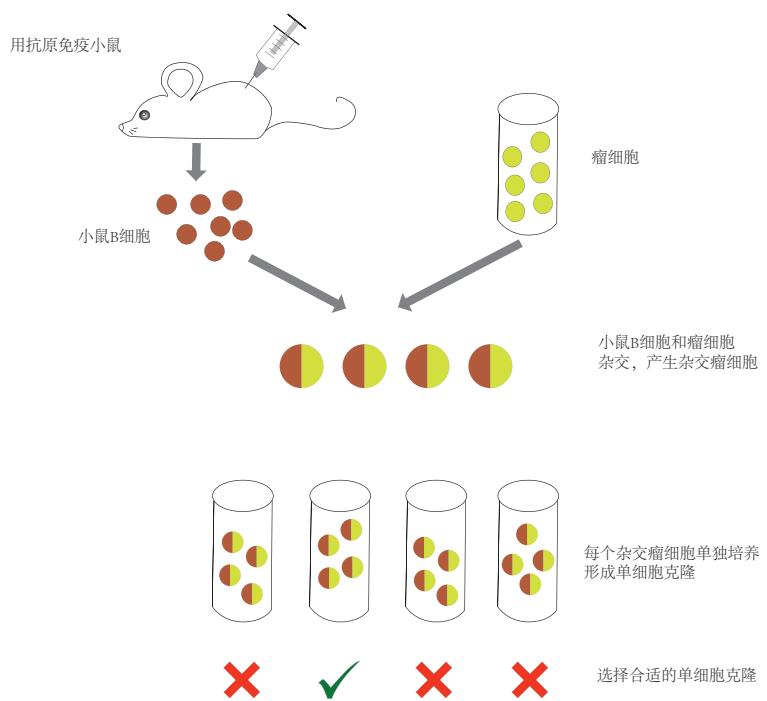


图 1.5: 单克隆抗体制备示意图

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。传统的方法产生有效的 CDR 需要大量的人力物力，如果可以计算的方法来精准预测这些 CDR，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的研究，可以指导 CDR 区域的设计，使得抗体可以针对指定的位点，并且具有足够高的亲和力，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。有的研究表面，重链上 CDR3 区域的多肽可以起到中和抗原的作用[6]。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，对抗原抗体相互作用的研究也可以为更广泛的蛋白质与蛋白质相互作用的研究提供借鉴。

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，并不能对抗原抗体的相互作用做出有效回答。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)是抗原抗体相互作用的关键[7]；Yokota 等通过对 HyHEL-10-HEL 复合物中一些氨基酸的突变(L-Y50F, L-S91A, L-S93A)来研究氢键在抗原抗体相互作用中的角色[8]，后来又研究了 Arg 在抗原抗体复合物中的作用[9]；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[10]；Shiroishi 等通过把 Tyr 突变成 Phe 和 Ala 来研究 Tyr 在 HyHEL-10-HEL 中的作用[11]；同样，Shiroishi 等通过对 HyHEL-10-HEL 的分析，研究了盐桥(salt-bridge)的作用[12]；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[13]。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[14, 15, 16, 17, 18, 19, 20, 21, 22]。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，都是关于某一个特定的复合物中特定氨基酸的研究，或者某种特定相互作用的研究，从来没有一个系统的关于所有抗原抗体相互作用的描述。其中一个重要原因，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）。

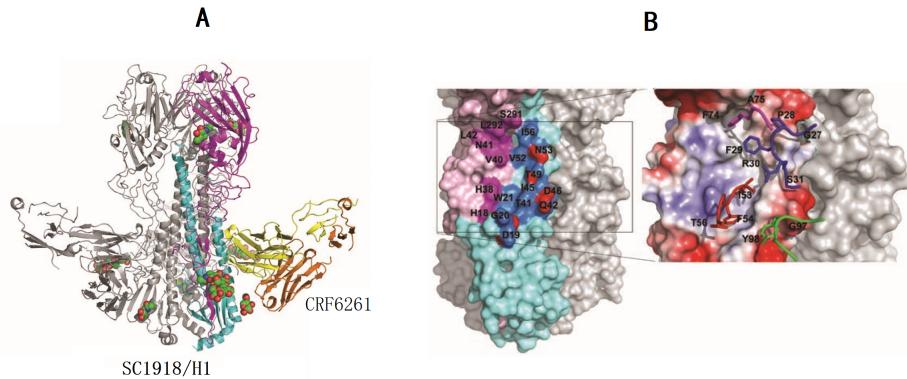


图 1.6: A 是 SC1918/H1 和 CR6261 复合物的结构。B 是 SC1918/H1 和抗体结合区域的放大图，其抗原表位都已经标出。此图由文献[23]中的图片编辑而成。

从图 1.6 B 中可以看出，在序列上，这些参与抗原抗体相互作用的氨基酸在序列并不连续，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 个抗原表位的研究，只有大概 4% 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[24]。这些结果说明，抗原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[25]。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[26]。虽然这篇文章发表在 2008 年，但却是在这之前以及这之后的许多抗原表位预测方法的理论依据，这些方法都是对任意给定的抗原来预测它的抗原表位。例如，Hopp 等通过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[27]； Rubinstein 等通过对已知的抗原表位的分析，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[29]。Xu 等对 CEP[30]，DiscoTope[31]，PEPOP[32]，ElliPro[33]、BEpro[34] 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，远不能满足实际需要[35]。

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[36]。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[37]。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 种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 与 x 相互作用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 抗体特异的抗原表位倾向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 α 碳与 i 的 α 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。在上面的公式种, $f_{ep}(xy)$ 可以算作对抗原抗体相互作用的一种描述, 但是这种描述只是计算特定氨基酸对出现频率的大小。

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[40]。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[41, 42]):

- (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[42]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。 χ 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。步骤(2)和步骤(3)的是作对相互作用的氨基酸对，或者相互作用的氨基酸三角（三个相互作用的氨基酸组成的三角形）的描述。这些描述可以看作是对抗原抗体相互作用的一种刻画，但是仍然局限于计算这些氨基酸对或者氨基酸三角的频率。

Liang Zhao 和 Jinyan Li 设计了Bepar(B-cell epitope prediction through association rules)[43]。其大概方法如图1.7：

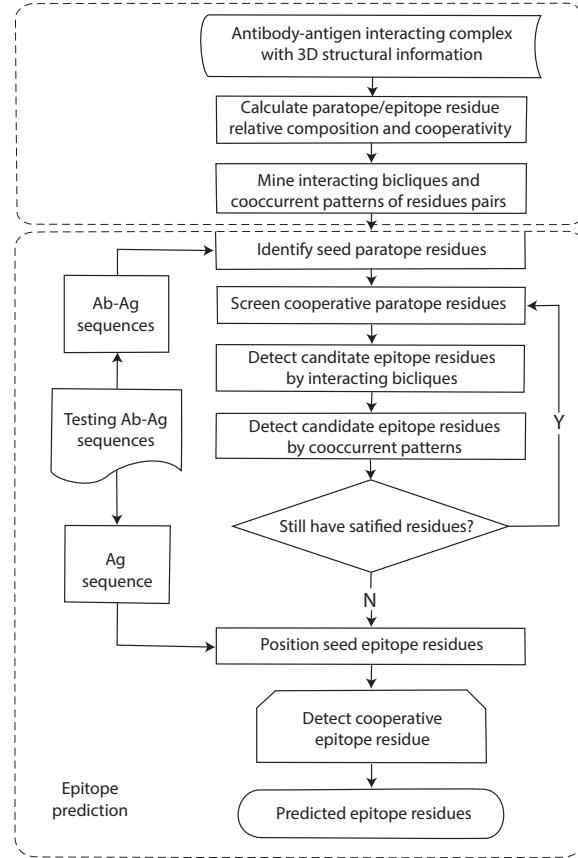


图 1.7: 此图源自文献[43]

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) 相关法则 (Association Rule) 的计算[44]。假设 A 和 B 分别是一些相互作用的氨基酸对的集合，并且 $A \cap B = \emptyset$ 。support A 是包含 A 的复合物的比例。相关法则 $R : A \rightarrow B$ 是指当观测到 A 就可以观测到 B 。support R 是指包含 $A \cup B$ 的复合物的比例。 R 的置信水平 (confidence level) 定义为：

$$\frac{\text{support } R}{\text{support } A}$$

作者选取 support $R \geq 10\%$ 并且置信水平 $\geq 80\%$ 的相关法则。

在计算过程中，Relative composition 被用来寻找种子，然后通过 Cooperativity 和相关规则进行扩展。同时，这些步骤也可以看成对抗原抗体相互作用的一种描述，可是这种描述仍然可以该概括为对相互作用的氨基酸对的数目的计算。

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[39]。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（每条抗体链上有3个 CDR 区域）。
- (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:

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

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。

PEASE 所构建的 Random Forest 把一个氨基酸对的 10 个特征联系了起来, 可以看成对抗原抗体相互作用的一种描述。但是, 这 10 个特征很少涉及氨基酸的生物化学性质, 同时通过 Random Forest 也是一种对氨基酸对的计数。

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, 但是里面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, 也就是计数。并没有运用氨基酸与氨基酸之间的替换矩阵(substitution matrix)来描述氨基酸的生化性质, 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, 可以描述抗原抗体相互作用的模型。这样也就基本确定了本论文的核心亮点: 提取相互作用的氨基酸序列对, 定义距离, 构建模型。

正如之前所说的, 抗原抗体相互作用的研究之所以如此困难, 关键就在于这种相互作用是一种三维结构的相互作用, 很难通过分析连续的序列来寻找规律。但是从另一个角度讲, 抗原抗体的相互作用可以看成一些相互作用的连续的氨基酸序列之间的组合, 只是这些连续的氨基酸序列的长度可能比较短, 甚至只含有一个氨基酸。之前的研究表明, 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)[45]。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[46]。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[48]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[49, 50]。

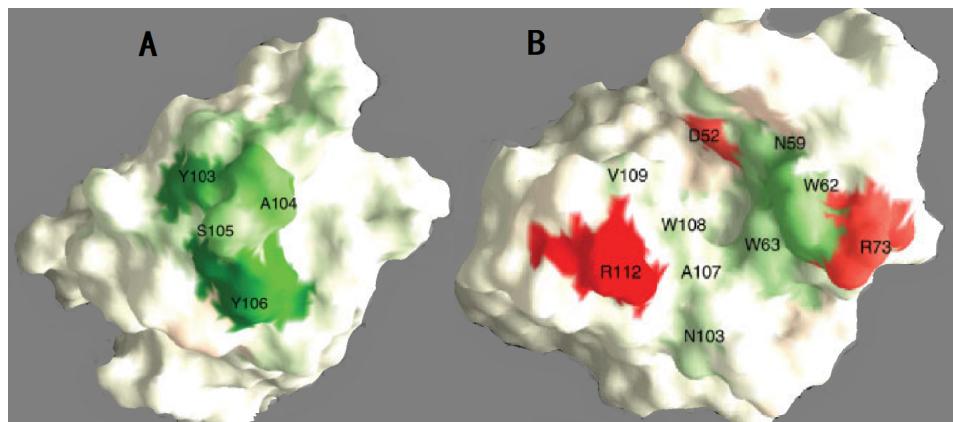


图 1.8: A 为抗体 cAB-Lys3 和上面的热点; B 为抗原 HEL(蛋清溶菌酶)上的热点。图片来自[50], 绿色代表加强抗原抗体的相互作用, 红色代表减弱抗原抗体的相互作用。

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, 为了排除那些并不重要的氨基酸序列对的干扰, 我们可以从每个抗原抗体复合物中选取最重要的相互作用的氨基酸序列对进行分析。在本论文中, 这些重要的相互作用的氨基酸对称之为 Core。

第3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[61]。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[62]，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)。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\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 \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。

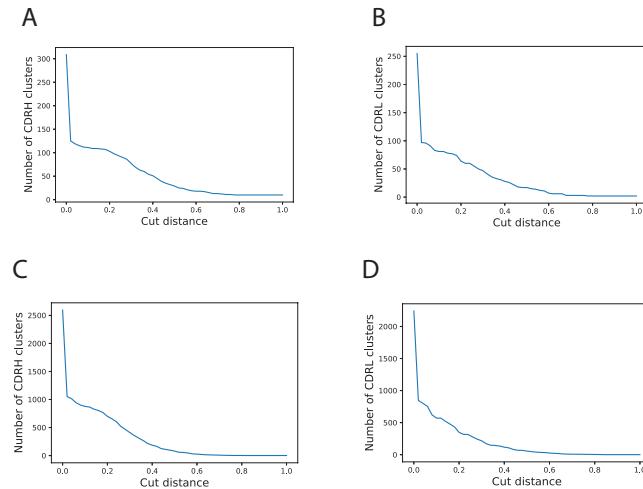


图 2.2: 截断距离和聚类数之间的关系。A 为 testing set 中重链的聚类情况，B 为 testing set 中轻链的聚类情况，C 为 training set 中重链的聚类情况，D 为 training set 中轻链的聚类情况。

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)的有: ([101], [13, 14, 15], 8)。
 - (d) 匹配类型为(2,1)的有: ([99, 100], [196], 4), ([101, 102], [196], 13)。
 - (e) 匹配类型为(2,2)没有。在这里要注意一个顺序问题，假设101 和 197 的CN数是10，那么 101, 102和 196, 197 的匹配是写成([101, 102], [196, 197], 26)，还是([101, 102], [197, 196], 26)呢？本论文在确定氨基酸顺序的时候，抗体CDR上氨基酸的顺序，一律从编号小的到编号大的，而抗原氨基酸的顺序由抗体氨基酸序列两端的氨基酸和抗原氨基酸两端的氨基酸相互作用情况而定。在这里，抗体101位氨基酸和抗原197位氨基酸的CN为10，101和196为8，102和197为3，102和196为5，因为 $8 + 3 < 10 + 5$,也就是把196 和197的顺序颠倒过来和有利，所以应该写成([101, 102], [197, 196], 26)。
 - (f) 匹配类型为(2,3)没有。
 - (g) 匹配类型为(3,1)的有([100, 101, 102], [196], 16)。
 - (h) 匹配类型为(3,2)的没有。
 - (i) 匹配类型为(3,3)的没有。
- (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。本论文对不同的匹配类型

都抽取了关键的相互作用的氨基酸序列，称为Core，并假设在同一个匹配类型下同一条链上最多有一个Core。所以表2.3最终Core的抽取结果如表2.4：

Match-type	Core
(1,1)	([52], [16], 9)
(1,2)	([30], [16, 17], 14)
(1,3)	([101], [13, 14, 15], 8)
(3,1)	([100, 101, 102], [196], 16)

表 2.4: 表2.3中最终 Core 的抽取结果。

第 2 节 数据的初步分析

从抗原抗体复合物中抽取出Core之后，对这次额Core的中氨基酸简单的统计分析就可以得出很多有意思结论。有些结论已经被以往的研究证实，有些则是全新的，还有些是在原有结论的基础上，有了进一步的发展。

(1) Core 的分布

我们从上面数据提取的过程，可以察觉到随着抗原氨基酸序列长度或者抗体氨基酸序列长度的增加，可以被抽取的Core的数目越来越少，这也符合一个简单的逻辑事实，那就是如果可以抽取较长的氨基酸序列的Core，那么就一定可以抽取较短的氨基酸序列的Core(表2.5)。

		Length of Antigen Amino Acids		
		1	2	3
Length of Antibody Amino Acids	1	1577	1432	934
	2	1454	1316	915
	3	1219	1130	805

表 2.5: 不同 match-type 下 Core 的数目。每行代表不同的抗体氨基酸的长度，每列代表不同的抗原氨基酸的长度。比如，match-type 为 (2, 1) 时 Core 的数目是1454。

为了更为直观的表示 Core 的数目随着氨基酸长度的变化，可以把表2.5的数据画成三维坐标图2.3。

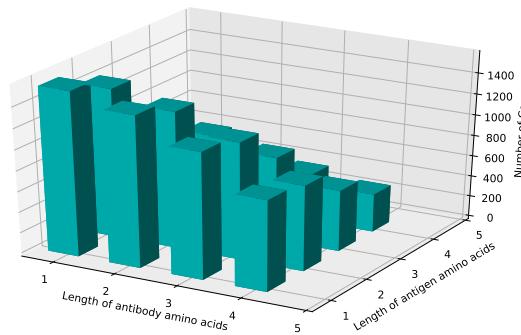


图 2.3: Core 的数目按不同 match-type 的分布

从图2.3 可以明显的看出，随着氨基酸序列长度的增加，Core 的数目在减少。

过去的研究表明重链上的CDR3区域在抗原抗体识别的过程中往往扮演着比其他的CDR更为重要的角色，有时甚至于决定性的作用，那么重链上的CDR3也应该比其他的CDR抽取更多的Core。

	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
CDRH1	203	189	132	173	149	112	157	152	109
CDRH2	328	261	153	253	213	127	229	191	109
CDRH3	424	454	343	487	490	387	428	418	341
CDRL1	280	198	107	174	137	74	140	115	61
CDRL2	84	59	32	61	44	24	27	23	14
CDRL3	258	271	167	306	283	192	238	231	171

表 2.6: 不同 match-type 的 Core 在不同CDR上的分布

从表2.6中的数据也可以看出，在绝对数量上，重链上CDR3占有明显的多数。如果我们把重链和轻链的数据分开，它们的CDR3都占有明显优势。从提取CDR的范围上说，CDR3 比 CDR1 和 CDR2 都要长，尽管本论文认为不是长度造成的，为了严谨，有必要在 Normalize 之后再比较（表2.7）。

	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
CDRH3	0.944	0.025	6.14E-5	5.74E-5	3.51E-10	4.84E-14	6.79E-4	5.44E-6	1.49E-11
CDRL3	0.226	5.41E-8	9.73E-8	1.89E-15	0	0	6.3E-15	0	0

表 2.7: 表中列出的是对不同的 match-type 情况下做假设检验, 得到的P值。对于给定的 match-type 和抗体链 (轻链或者重链), H_0 : CDR3 上 Core 的密度大于其他两个CDR。这里密度, 是用Core 的数目除以 CDR 的长度。

从表2.7可以看出, 除了 match-type 为 (1, 1) 的情况外, 其他 match-type 下, CDR3 上 Core 的密度明显大于其他 CDR。这从另外一个角度验证了 CDR3 在抗原抗体作用过程中的重要性。

- (2) Core 中氨基酸的分布 如果观察所有 Core 中抗原氨基酸和抗体氨基酸的分布 (图2.4), 会发现有明显差异。

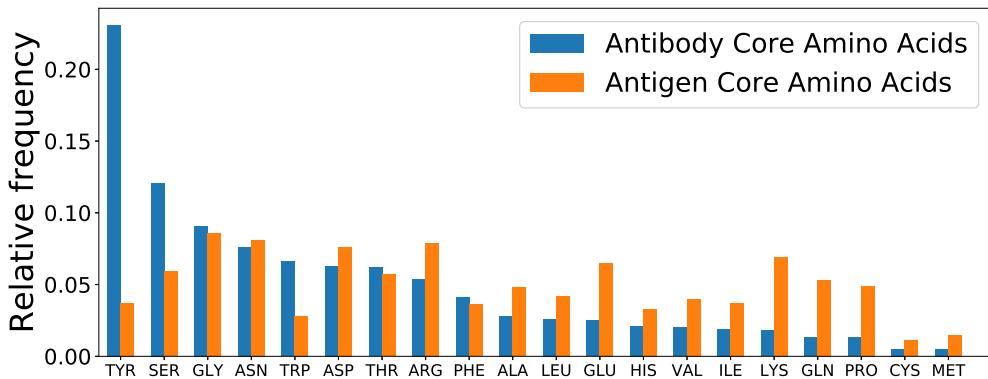


图 2.4: 抗原氨基酸和抗体氨基酸在所有Core中的分布情况

从图2.4中可以发现, TYR 在 Core 中抗体氨基酸的分布明显偏高, 所占比例竟然超出了20%! 对于 match-type 为 (1, 1) 的 Core 来讲, 频率最高的前8个分别为 ([TYR], [ARG]), ([TYR], [LYS]), ([TYR], [HIS]), ([TYR], [ASN]), ([TYR], [GLU]), ([TYR], [PRO]), ([TYR], [ASP]), ([TYR], [GLN]), 在抗体氨基酸竟然全部是TYR! 如果把不同 match-type 下的抗体氨基酸的分布分别来考虑, 从表2.8可以看出, 在任何 match-type 的情况下, TYR 的比例在 Core 的抗体氨基酸中分布比例都是最高的。文献[63, 64] 对 TYR 的重要性也有过描述。其中一个解释, 是说 TYR 的侧链有一个羟基苯, 苯环的疏水作用, 使得羟基形成的氢键非常牢固。

Match-type	Top five most frequent antibody amino acids				
(1,1)	(TYR, 0.361)	(TRP, 0.120)	(ARG, 0.091)	(SER, 0.073)	(ASN, 0.068)
(1,2)	(TYR, 0.357)	(TRP, 0.120)	(ASN, 0.074)	(ARG, 0.071)	(SER, 0.067)
(1,3)	(TYR, 0.336)	(TRP, 0.108)	(ASN, 0.096)	(ARG, 0.077)	(SER, 0.07)
(2,1)	(TYR, 0.208)	(SER, 0.142)	(GLY, 0.102)	(ASN, 0.079)	(THR, 0.070)
(2,2)	(TYR, 0.237)	(SER, 0.123)	(GLY, 0.100)	(ASN, 0.071)	(THR, 0.066)
(2,3)	(TYR, 0.234)	(SER, 0.124)	(ASN, 0.090)	(GLY, 0.090)	(TRP, 0.060)
(3,1)	(TYR, 0.176)	(SER, 0.138)	(GLY, 0.115)	(ASN, 0.074)	(THR, 0.074)
(3,2)	(TYR, 0.190)	(SER, 0.136)	(GLY, 0.112)	(THR, 0.073)	(ASP, 0.070)
(3,3)	(TYR, 0.193)	(SER, 0.128)	(GLY, 0.111)	(ASN, 0.075)	(THR, 0.065)

表 2.8: 不同 match-type 下, 前五个分布频率最高的抗体氨基酸

如果仔细观察表格2.8, 我们发现当抗体氨基酸的序列的长度从1变成2或者3时候, 前三个频率最高的氨基酸由 TYR, TRP, ASN/ARG 变成 TYR, SER, GLY(match-type 是 (2, 3) 时候, GLY处于第四位, 但是和第三位的 GLN 在比例上很近)。GLY 和 SER 的侧链很小, 方便调节局部的构象, 所以可以推断 GLY 和 SER 协同 TYR 参与对抗原表位的识别。Sara Birtalan 等研究发现, 只用 TYR, SER 和 GLY 这三种氨基酸就可以造出亲和力近于pM的抗体[65]。这个结论在表2.8中得到充分的支持。Frederic A. Fellouse 等用 TYR, SER, GLY 和 ASP 四种氨基酸构建抗体 library, 成功筛选出抗 hVEGF 的特异性抗体[66], 并指出了TYR的主导作用[67]。表2.8 也为这些结论提供了充分的支持。在为已有的结论提供支持的同时, 也可以指导新的抗体 Library 的建立。根据上面的表格, 如果想用四个氨基酸来建设抗体 Library, 显然最好的选择是 TYR, SER, GLY 和 ASN, 而不是Frederic A. Fellouse 等使用的TYR, SER, GLY 和 ASP。

表2.9给出了 Core 中抗原表位氨基酸在不同 match-type 下的分布。可以看出来, 在所有的 match-type 情况下, ARG 和 ASP 的出现频率一直位于 Core 中抗原氨基酸的前5。而且当抗原氨基酸的长度由1变成2或3的时候, GLY忽然从前5名之外变成出现频率最高的氨基酸 (在match-type(2,3)下, GLY 位于第二的位置, 但是和出现频率最高的 ASN 相差只有0.001)。除了 GLY 之外, 其他排在前5的氨基酸, 都是亲水, 而且多由比较长的侧链。由此可以说, 在 GLY 的调节下, 具有比较大的亲水的氨基酸往往可以称为抗原表位, 特别是 ARG 和 ASP。

Match-type		Top five most frequent antigen amino acids				
(1,1)	(ARG, 0.139)	(ASP, 0.091)	(LYS, 0.083)	(GLU, 0.083)	(ASN, 0.078)	
(1,2)	(ARG, 0.131)	(ASP, 0.094)	(GLU, 0.094)	(ASN, 0.087)	(LYS, 0.085)	
(1,3)	(ARG, 0.130)	(GLU, 0.092)	(ASP, 0.092)	(LYS, 0.089)	(GLN, 0.084)	
(2,1)	(GLY, 0.097)	(ASP, 0.076)	(ASN, 0.073)	(ARG, 0.070)	(LYS, 0.069)	
(2,2)	(GLY, 0.088)	(ASN, 0.081)	(ASP, 0.076)	(ARG, 0.071)	(LYS, 0.069)	
(2,3)	(ASN, 0.080)	(GLY, 0.079)	(ASP, 0.075)	(LYS, 0.074)	(GLU, 0.069)	
(3,1)	(GLY, 0.108)	(ASN, 0.082)	(SER, 0.067)	(THR, 0.065)	(ALA, 0.064)	
(3,2)	(GLY, 0.103)	(ASN, 0.084)	(ASP, 0.068)	(THR, 0.068)	(SER, 0.067)	
(3,3)	(GLY, 0.105)	(ASN, 0.083)	(SER, 0.070)	(ASP, 0.067)	(THR, 0.066)	

表 2.9: 不同 match-type 下, 前五个分布频率最高的抗原氨基酸

为了更进一步说明 Core 中抗原氨基酸的分布情况, 本论文把 Core 中频率最高的5个抗原氨基酸和频率最低的5个氨基酸, 和它们在所有抗原氨基酸中的分布情况做了对比, 看看它们这些分布的富集和稀释是不是显著(表??)。从表2.10中也可以看出被富集的氨基酸除了GLY之外都是亲水性的氨基酸, 特别是ARG, GLY则帮助这些亲水性的氨基酸形成合适的结构。从表2.11中可以看到, 在所有的 match-type 中, VAL, LEU, ILE 和 CYS 位于被稀释的前五名之列。CYS容易被被氧化, 形成二硫键, 所以不适合做抗原表位。其他三个都是疏水性的氨基酸。观察表2.11, 几乎所有的都是疏水性的氨基酸(只有在 match-type(3,2)的第五名出现SER)。这些结果说明, 亲水性有比较大的侧链的氨基酸容易形成抗原表位, 而疏水性的不容易形成。这和普通的蛋白质和蛋白质之间的相互作用不同[68], 同时也解释了水分子在抗原抗体相互作用过程中的重要性[69]。

Match-type		Top 5 enriched residues in the Core Antigen residues				
(1,1)	(ARG, 0.0E+00)	(HIS, 0.0E+00)	(TRP, 0.0E+00)	(ASP, 1.6E-12)	(GLN, 1.4E-07)	
(2,1)	(ARG, 0.0E+00)	(TRP, 6.8E-14)	(ASP, 3.8E-13)	(GLU, 3.3E-09)	(HIS, 5.3E-09)	
(3,1)	(ARG, 0.0E+00)	(GLN, 2.3E-13)	(TRP, 5.7E-12)	(ASP, 2.2E-10)	(GLU, 3.0E-07)	
(1,2)	(ARG, 1.5E-09)	(ASP, 7.6E-09)	(GLY, 6.1E-08)	(HIS, 3.4E-05)	(ASN, 4.8E-04)	
(2,2)	(ARG, 1.6E-09)	(ASP, 3.4E-08)	(ASN, 4.3E-07)	(GLY, 7.9E-04)	(HIS, 1.0E-03)	
(3,2)	(ARG, 1.6E-06)	(ASP, 1.2E-06)	(ASN, 9.0E-06)	(TRP, 9.6E-06)	(GLN, 2.7E-04)	
(1,3)	(GLY, 4.9E-14)	(ASN, 7.0E-08)	(ARG, 5.3E-05)	(ASP, 4.7E-03)	(PRO, 1.5E-02)	
(2,3)	(GLY, 1.6E-10)	(ASN, 6.7E-09)	(ASP, 1.9E-04)	(ARG, 1.1E-03)	(HIS, 2.6E-02)	
(3,3)	(GLY, 2.0E-10)	(ASN, 2.1E-07)	(ASP, 1.3E-03)	(ARG, 5.4E-03)	(LYS, 4.1E-01)	

表 2.10: 和所有抗原氨基酸相比, 富集最为显著的抗原氨基酸(按照q value 排序)。富集是指和所有抗原氨基酸的分布相比, 相对频率增高的氨基酸。

Match-type		Top 5 diluted residues in the Core Antigen residues				
(1,1)	(VAL, 1.5E-15)	(ALA, 5.2E-14)	(LEU, 5.0E-12)	(ILE, 1.0E-11)	(CYS, 1.5E-08))	
(2,1)	(LEU, 8.5E-19)	(VAL, 1.3E-12)	(ALA, 1.1E-10)	(ILE, 2.5E-10)	(CYS, 2.1E-07)	
(3,1)	(LEU, 2.6E-16)	(VAL, 6.9E-10)	(ILE, 1.1E-07)	(ALA, 1.3E-07)	(CYS, 2.3E-06)	
(1,2)	(LEU, 9.7E-10)	(CYS, 3.8E-09)	(VAL, 9.6E-08)	(ILE, 5.0E-06)	(ALA, 1.8E-03)	
(2,2)	(LEU, 1.1E-11)	(VAL, 1.1E-07)	(CYS, 1.1E-07)	(ILE, 9.3E-05)	(ALA, 2.4E-03)	
(3,2)	(LEU, 1.3E-10)	(CYS, 5.0E-06)	(ILE, 7.9E-05)	(VAL, 1.7E-04)	(SER, 5.9E-03)	
(1,3)	(LEU, 1.2E-13)	(CYS, 1.1E-04)	(VAL, 1.2E-04)	(ILE, 1.4E-04)	(PHE, 1.7E-02)	
(2,3)	(LEU, 3.6E-12)	(VAL, 8.5E-04)	(ILE, 4.8E-03)	(CYS, 5.1E-03)	(PHE, 8.4E-03)	
(3,3)	(LEU, 5.4E-13)	(CYS, 1.8E-03)	(ILE, 9.0E-03)	(VAL, 1.1E-02)	(PHE, 5.6E-02)	

表 2.11: 和所有抗原氨基酸相比，释希最为显著的抗原氨基酸（按照q value 排序）。这里稀释是和所有抗原氨基酸的分布相比，相对频率下降的氨基酸。

在所有的相互作用的抗原氨基酸和抗体氨基酸对中，如果两个参与相互作用的氨基酸，中间间隔一个氨基酸，我们把这样的氨基酸称为中间氨基酸(middle amino acid)。把所有的中间氨基酸都提取出来进行比较，可以画出图2.5。通过分析发现，抗体上的中间氨基酸前频率最高的前5名分别是GLY, PRO, ILE, THR 和 SER，抗原上的氨基酸则是 LEU, ILE, VAL, ALA, GLY。前面的分析结果表明 LEU, ILE, VAL 是抗原表位所排斥的，所以，很有可能这些疏水性的氨基酸就像一个锚一样插入到抗原的内部，把两侧可以参与抗原抗体识别的氨基酸锚定在表面。在抗原的中间氨基酸中，PRO 占有比较大的比例，PRO 有形成转角的作用，很有可能协调它两侧的氨基酸共同参与对抗原的识别。

The Middle Amino Acids

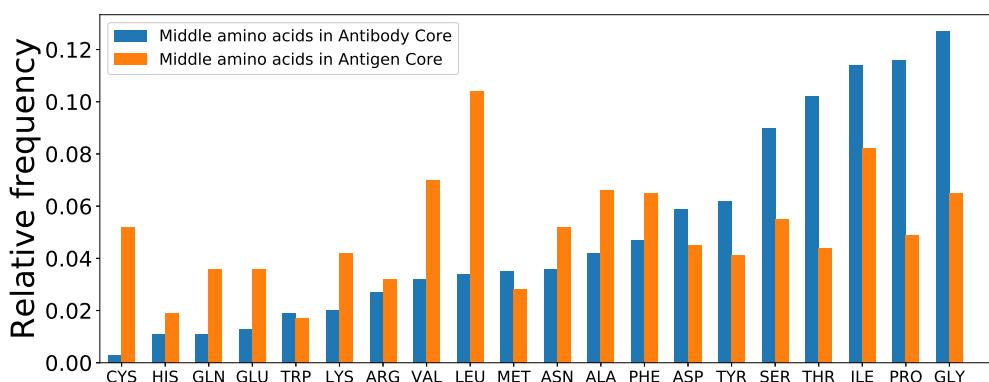


图 2.5: 中间氨基酸的组成比例

第三章 RBFN 的构建

Dash 等[70] 用10重不同的抗原免疫产生了4600多个 $\alpha\beta$ TCR, 通过对这些 $\alpha\beta$ TCR 的分析, 作者数量化的刻画了抗原和 $\alpha\beta$ TCR之间的关系。但是到据我们所知, 到目前还没有类似的刻画抗原抗体作用的模型, 本章内容就是区建立这样一个模型。

第 1 节 Core 之间距离的定义

径向基函数网络(RBFN, radial basis function network)要建立在径向基函数上, 而径向基函数的输入值是一个距离。其含义就是, 对于一个选定的center, 在径向基函数下, 和它距离相同的点具有相同的性质, 或者说这个 center 对它周围的影响只于距离有关。要建立 RBFN , 第一步就是要定义距离, 定义的好坏直接关系到RBFN的好坏。Dash 等[70] 在建立模型时候用到BLOSUM62, 但是却是一个截断的BLOSUM62, 基本规则如下:

- (1) $AAdist(a, a)=0$
- (2) $AAdist(a, b)=\min(4, 4-BLOSSUM62(a,b))$
- (3) $AAdist(a, -)= 4$, if a is in CDR1 or CDR2
- (4) $AAdist(a, -) = 8$, if a is in CDR3

在上面的规则中, $AAdist$ 表示氨基酸和氨基酸之间的距离。- 表示空缺。上面的定义有两个不足。一是, 上面的计算距离的时候是在 align 之后, 然而 align 时候本身就有自己的打分规则, 现在又再次进行打分, 逻辑上不可行, 不如直接用上面的规则进行align。二是, 对 BLOSUM62 进行截断的理由不足。BLOSUM62 是通过大量同源氨基酸之间的比较得到的一个打分矩阵, 反应的是氨基酸之间差异的一个综合分数, 这个矩阵的最大值是 TRP 和 TRP 之间的相似分数 11, 最小值是 CYS 和 GLU 之间的相似分数 -4。如果用4 对BLOSUM62 进行截断, 可能会使得本该反应出来的差异得不到充分反应。所以本论文没有对BLOSUM62进行截断。我们模型的功能是给定对短的抗体氨基酸的序列和短的抗原氨基酸的序列, 来判定它们是否可以称为 Core, 也就是说是否能在抗原抗体的相互作用中扮演重要作用。所以我们定义短序列对和短序列对之间的距离。设 $(Ab1, Ag1)$ 和 $(Ab2, Ag2)$ 为 match-type (m, n) 的短序列对, 那么它们之间的距离可以通过下面的步骤来定义:

(1) 定义 Ab1 和 Ab2 之间的相似度，平移成非负值，并 Normalize，如下：

$$S_{Ab} = Aln(Ab_1, Ab_2) \quad (3.1)$$

$$S_{Ab}^+ = \frac{S_{Ab} + 4 \times m}{15 \times m} \quad (3.2)$$

在公式(3.1)中，Aln 表是序列比对 (sequence alignment)的分数，序列比对所用的规则是：

Substitution matrix = BLOSUM62

gap = Hp_1

extended gap = Hp_2

在上面的规则中，我们完全采用了 BLOSUM62； Hp_1 和 Hp_2 表示两个超参数 (hyperparameter)，它们的取值范围是 $\{-1, -2, -3, -4, -5\}$ 。在公式 (3.2) 中， $4 \times m$ 和 $15 \times m$ 来自于 Substitution Matrix 中最小的值是-4，最大值是11.

(2) 以同样的方式定义 Ag1 和 Ag2 之间的相似度，平移成非负，并 Normalize，得到

S_{Ag}^+

(3) 定义 (Ab_1, Ag_1) 和 (Ab_2, Ag_2) 之间的相似面积：

$$S = S_{Ab}^+ \times S_{Ag}^+ \quad (3.3)$$

这里我们计算面积，而没有计算长度，是因为 S_{Ab}^+ 和 S_{Ag}^+ 是协同作用的，而不是独立作用的。这个论断来自一个简单的事实，当 (Ab_1, Ag_1) 有强烈的相互作用，并且 Ag_2 和 Ag_1 完全相反时， (Ab_1, Ag_2) 之间的相互作用应该是0。也就是说， Ab_1 和 Ab_2 之间的相似程度所起到的作用和 Ag_1 和 Ag_2 之间的相似性相关。反过来也一样。

(4) 计算 (Ab_1, Ag_1) 和 (Ab_2, Ag_2) 之间的距离：

$$D = 1 - S \quad (3.4)$$

第 2 节 RBFN 模型

从抗原抗体复合物中提取出来的 Core 是具有相互作用的正样本。每种 match-type 下的负样本是随机产生的并且不在正样本中的氨基酸序列对。本文在每种 match-type

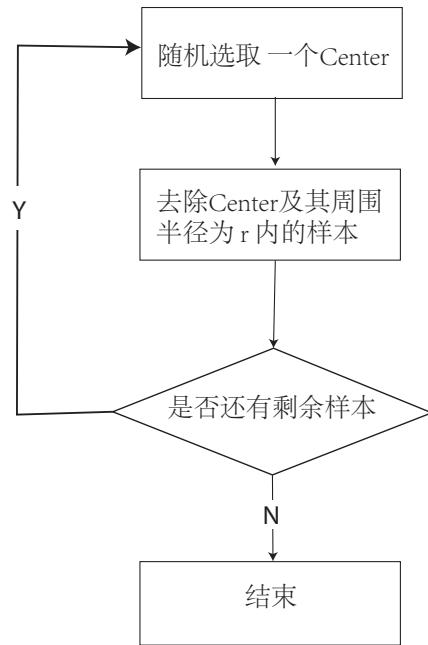


图 3.1: 文献[71]选取 center 的流程图

下独立的产生10组负样本，每组负样本的样本量和正样本一样。

RBFN 的径向基函数选择为高斯函数，对于 center c 来说可以表示为如下：

$$f(x, r, c) = \exp\{-d(x, c)^2/r^2\}$$

RBFN 可以表示为：

$$\text{RBFN}(x) = \sum w_i f(x, r_i, c_i)$$

其中 w_i 和 r_i 是要在模型训练过程中确定的参数。对于 RBFN 的 center 的选取，本论文采用 Alexandridis 等在文献[71] 中提供的方法。具体步骤可以概括流程图3.1：

从本质上讲，上面的 center 选择方法是聚类的方法，首先对样本进行聚类，然后在每个类中任意选出一个代表。本论文就是采用的聚类方法，这样做不但可以根据半径的大小来确定 center，而且可以根据所要选择 center 数目的多少来确定 center。

本论文尝试了两种观测值，一种是二分类观测，也就是从复合物中提取出来的观测值都定为1，随机产生的负样本的观测值定为1。第二种是把每个 Core 的 CN 数作为观测值，负样本的 CN 数记为 0. 对于二分类的观测值，损失函数选为 Sigmoid 函数并加上正则项，对于 CN 作为观测值，用方差作为损失函数并加上正则项。对于训练集，本论文采用 5-fold 的交叉检验(cross validation)。这里我们总结以下超参数都是哪些。

- (1) 在距离定义规则中 gap 的分数 Hp_1 , 取值范围是 $\{-1, -2, -3, -4, -5\}$
- (2) 在距离定义规则中 extended gap 的分数 Hp_2 , 取值范围是 $\{-1, -2, -3, -4, -5\}$

- (3) 正则项的系数 r , 取值范围是 (0.0001, 0.001, 0.01, 0.1)
- (4) Center 数目占训练集的比例 p , 取值范围是 0.8, 0.4, 0.2, 0.1, 0.05, 0.02
- (5) 观测值选 Binary 类型还是 Numerical 类型

表3.1 和表3.2 分别给出了在 CN 为观测值和二分类为观测值情况下, 交叉检验的最优表现和对应的超参数。可以看出, 二分类观测值在各种 match-type 下都优于 CN 观测值。本论文的模型确定为二分类模型, 超参数取值也都对应于二分类模型下的最优取值。

match-type	Hp_1	Hp_2	r	p	平均AUC
(1,1)	-1	-1	0.0001	0.8	0.955
(1,2)	-1	-1	0.001	0.8	0.842
(1,3)	-1	-1	0.001	0.8	0.808
(2,1)	-1	-1	0.001	0.4	0.851
(2,2)	-1	-1	0.01	0.8	0.825
(2,3)	-1	-1	0.01	0.8	0.812
(3,1)	-1	-1	0.01	0.8	0.844
(3,2)	-1	-1	0.01	0.8	0.839
(3,3)	-1	-1	0.01	0.8	0.842

表 3.1: 在观测值为 CN 时, 在各种 match-type 下的交叉检验的最有表现。这里的最优是按照交叉检验过程中平均AUC来衡量的。

match-type	Hp_1	Hp_2	r	p	平均AUC
(1,1)	-1	-1	0.0001	0.8	0.973
(1,2)	-1	-1	0.0001	0.8	0.860
(1,3)	-1	-1	0.0001	0.8	0.834
(2,1)	-1	-1	0.0001	0.8	0.870
(2,2)	-1	-1	0.0001	0.8	0.842
(2,3)	-1	-1	0.001	0.8	0.836
(3,1)	-1	-1	0.0001	0.8	0.867
(3,2)	-1	-1	0.001	0.8	0.862
(3,3)	-1	-1	0.001	0.8	0.862

表 3.2: 在观测值为 0, 1 时, 在各种 match-type 下的交叉检验的最有表现。这里的最优是按照交叉检验过程中平均AUC来衡量的。

鉴于样本量并不是特别大，在训练模型的时候选用的拟牛顿法（quasi-Newton）结合Ramijo 线搜索方法[72]。具体的程序和有关交叉检验和模型训练的结果请参看 https://github.com/Leochuanxing/Paritope_Epitope

检验结果 正如前面所说的，在提取数据的时候有10% 的抗原抗体复合物被留作正测试集，从中抽取的 Core 是正的测试集。负测试集的产生和负样本的生成采用同样的方法，随机生成并且不在正样本中的作为负测试集。在每一种 match-type 下，本论文独立的产生了 10 组负，且每组负测试集的大小和对应正测试集的大小一样。把正测试集分别和这10组负测试集混合产生 10 组测试集，然后用构建的 RBFN 模型进行区分，ROC 如图3.2

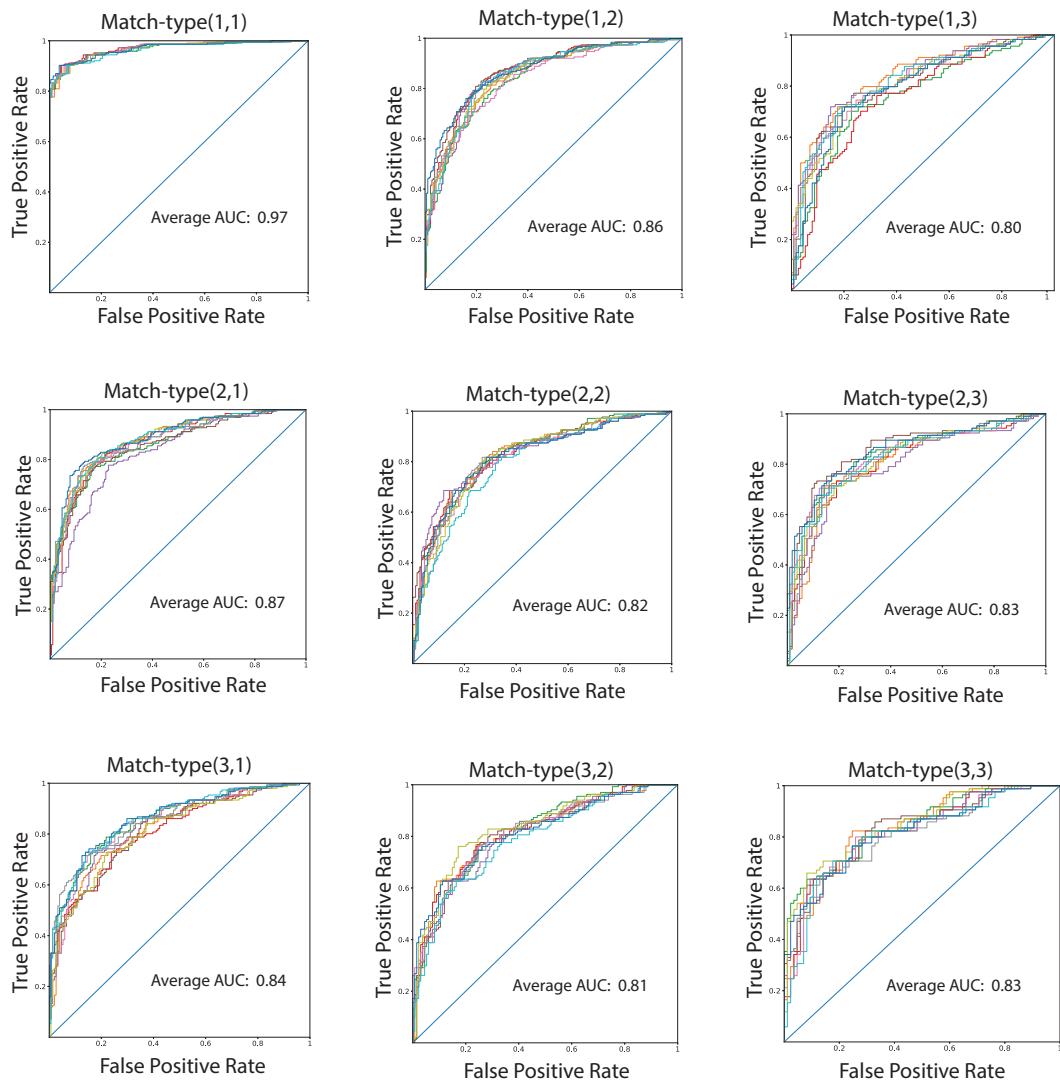


图 3.2: 在不同 match-type 下 RBFN 模型正负测试集的区分。在每种 match-type 下，有 10 条 ROC 曲线对应 10 组测试集，Average AUC 是这 10 个测试集 AUC 的平均。

从图3.2中可以看出，对于所有 match-type，平均 AUC 都不低于 0.8，在 match-

type 为 (1, 1) 时, 平均 AUC 高达 0.97! 也就是说, 任意给定一个氨基酸对和一个 match-type 为(1,1) 的 Core, 本论文的模型可以把二者区分开的概率是0.97。在 match-type 为 (1, 3) 时表现最差, 能够把真正的 Core 和随机产生的氨基酸序列对区分开的概率是0.8, 也已经比较高了。从这些结果中, 可以看出 RBFN 模型比较好的刻画了 Core 的性质, 也就是说比较好的刻画了抗原抗体相互作用过程中起关键作用的短的抗体氨基酸序列和短的抗原氨基酸序列之间的关系。当然更是可以得出, 抗原和抗体关键氨基酸之间的作用是有规律的这样的推论。

上面模型有一个让人诟病的地方, 就是对每一种 match-type 都要建立一个模型, 相当于9个模型。本文还做了一个把这些模型整合成一个模型的尝试。鉴于 match-type (1, 1) 情况下, 模型的表现如此优秀, 本文尝试以 match-type(1,1)的模型为基础来解释其他模型。假设 $([a_1, a_2], [b_1, b_2])$ 是一个 match-type(2,2) 的 Core, 那么这个 Core 可以组合成 4 个 match-type(1,1)的 Core.

$$([a_1], [b_1]) \quad ([a_1], [b_2]) \quad ([a_2], [b_1]) \quad ([a_2], [b_2])$$

设 $\text{RBFN}_{(1,1)}$ 为 match-type(1,1) 的 RBFN 模型, 那么 $([a_1, a_2], [b_1, b_2])$ 的模型返回值可以定义为

$$\frac{\sum_{i=1}^2 \sum_{j=1}^2 \text{RBFN}_{(1,1)}([a_i], [b_j])}{4}$$

对于其他 match-type 的 Core 也做如上面类似的定义。也就是说, 用match-type(1,1)的模型来解释所有 match-type 的模型, 本文称这个模型为整合模型。图3.3给出了整合模型的预测效果。

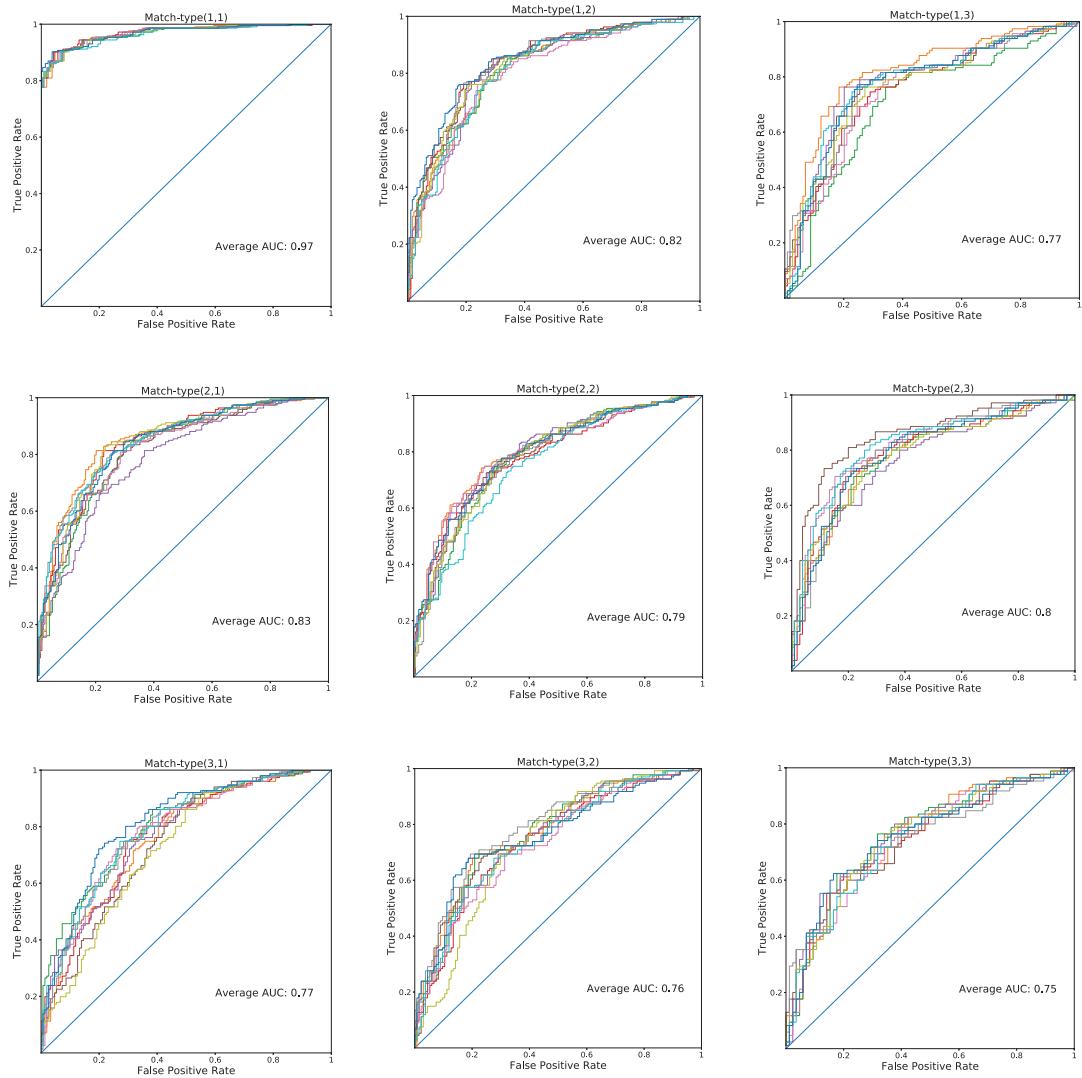


图 3.3: 在不同 match-type 下整合模型正负测试集的区分。在每种 match-type 下, 有 10 条 ROC 曲线对应 10 组测试集, Average AUC 是这 10 个测试集 AUC 的平均。

从图3.3可以看出, 整合模型和前面的RBFN模型在 match-type(1,1) 上的预测效果一样, 这也是显然的。但是, 对于其他的 match-type, 整合模型明显处于劣势, 平均 AUC 差距为 4.6%。最高的为 match-type(3,3), 差距达到 8%。这些差距是巨大的。而且似乎, 可以看出随着抗原氨基酸或者抗体氨基酸序列的增长, 整合模型表现的越来越差。

第四章 RBFN 模型的应用

前面一章建立了 RBFN 模型，并且这个模型在测试集上表现良好。同时还测试了一个整合模型，这个模型和RBFN模型相比，显得逊色。这一章将会把上面建立的模型用来解决两个问题，一个是区别 Core 的抗原氨基酸序列和抗体氨基酸序列，一个是预测突变对抗原抗体复合物的影响。在前面介绍本论文背景时，有不少篇幅是用来介绍抗原表位(epitope)预测的研究，其中原因就是因为抗原表位的预测需要对抗原抗体作用规律的研究，这也是本论文的核心。然而在这里，本论文并没有对把模型用于对抗原表位的预测。

第 1 节 Core 中抗原氨基酸和抗体氨基酸的区别

在前一章建立的模型中，输入值是两个短的氨基酸序列组成的对($AbSeq, AgSeq$)，其中 $AbSeq$ 是短的抗体氨基酸序列， $AgSeq$ 是短的抗原氨基酸序列。那么，从总体数据上来看， $AbSeq$ 和 $AgSeq$ 有没有区别呢？这就是本节内容所要解决的问题。

设($AbSeq, AgSeq$) 是 $match-type(m, n)$ 的 Core，如果 $AbSeq$ 和 $AgSeq$ 没有区别，那么 $(AgSeq, AbSeq)$ 和真正的 $match-type(n, m)$ 的 Core 没有区别，自然在前一章中构建的模型不能对二者做出有效区分。这里，对于一个 $match-type(m, n)$ 的测试集，它是由 $match-type(m, n)$ 的正测试集和 $match-type(n, m)$ 的正测试集通过交换抗原序列和抗体序列的前后关系得到。为表述清楚，我们用下面的数学语言表示 $match-type(n,m)$ 的测试集的生成。

$$TR_{(n,m)} = \{(AgSeq, AbSeq) : (AbSeq, AbSeq) \in T_{(m,n)}\}$$

$$T = TR_{(n,m)} \cup T_{(n,m)}$$

在上面的表述中， $T_{(m,n)}$ 和 $T_{(n,m)}$ 分别为前一章中 $match-type(m,n)$ 和 $match-type(n,m)$ 的测试集， $TR_{(n,m)}$ 为 $T_{(m,n)}$ 中的元素通过交换抗原和抗体氨基酸序列的前后顺序得到，简称为 TR (reverse testing set)。图4.1 给出了 RBFN 模型的区分效果。

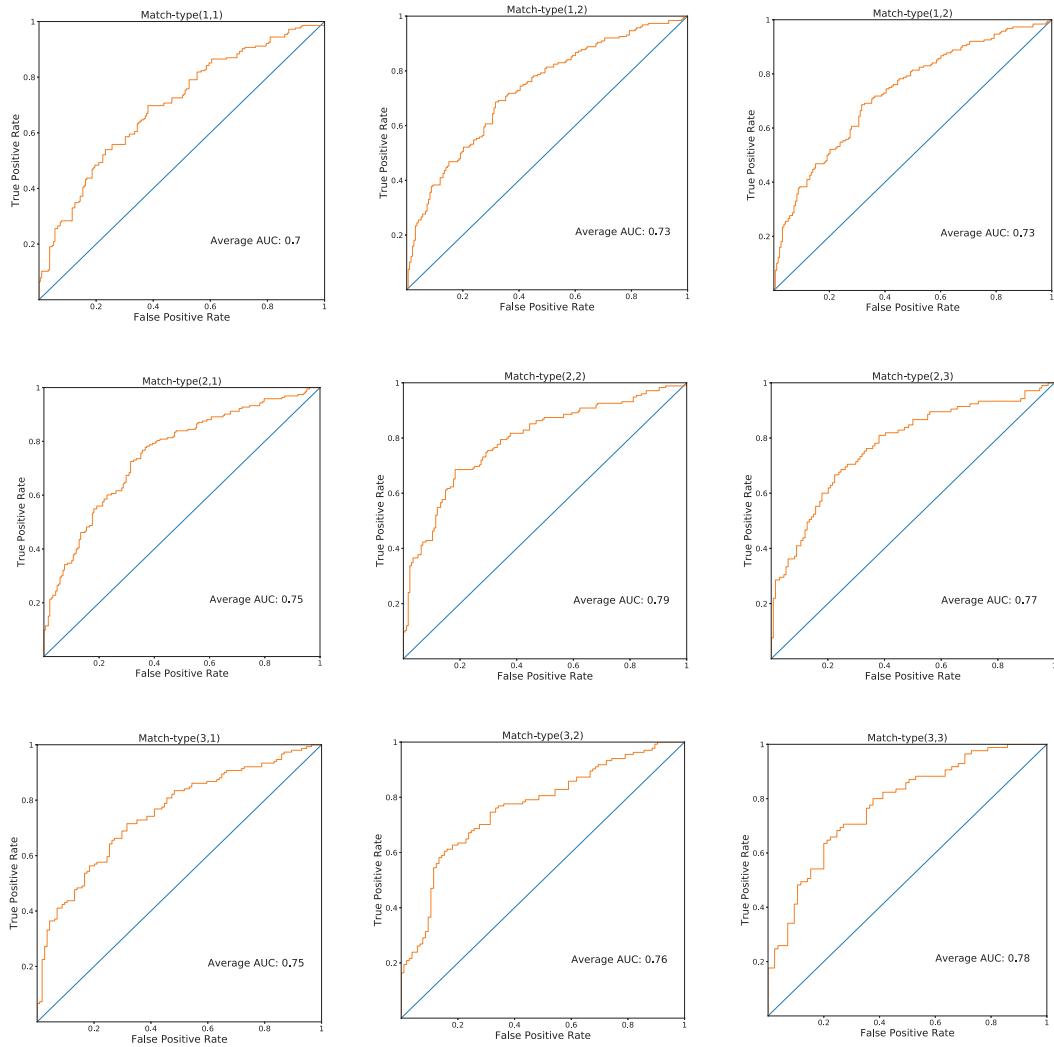


图 4.1: RBFN 用来区分测试集 T 和 TR 的ROC曲线和对应 AUC 值。

从图4.1可以看出，在不同的 match-type 下，RBFN 模型可以对 T 与 TR 做出较好的区分，对应 ROC 的 AUC 值从 0.7 到 0.78 不等。这个结果说明，Core 中的抗原氨基酸序列和抗体氨基酸序列有明显的区别。但是，这个区分效果弱于 T 与随机产生的负测试集之间的区分效果（图3.2），平均的AUC 差别为 9.6 %。这个差值可以理解为由 Core 中的抗原氨基酸序列和抗体氨基酸序列的相似性造成，也就是，从 AUC 的角度来讲，Core 中的抗原氨基酸序列和抗体氨基酸序列的相似性使得对它们的区分效果下降约9.6 %。

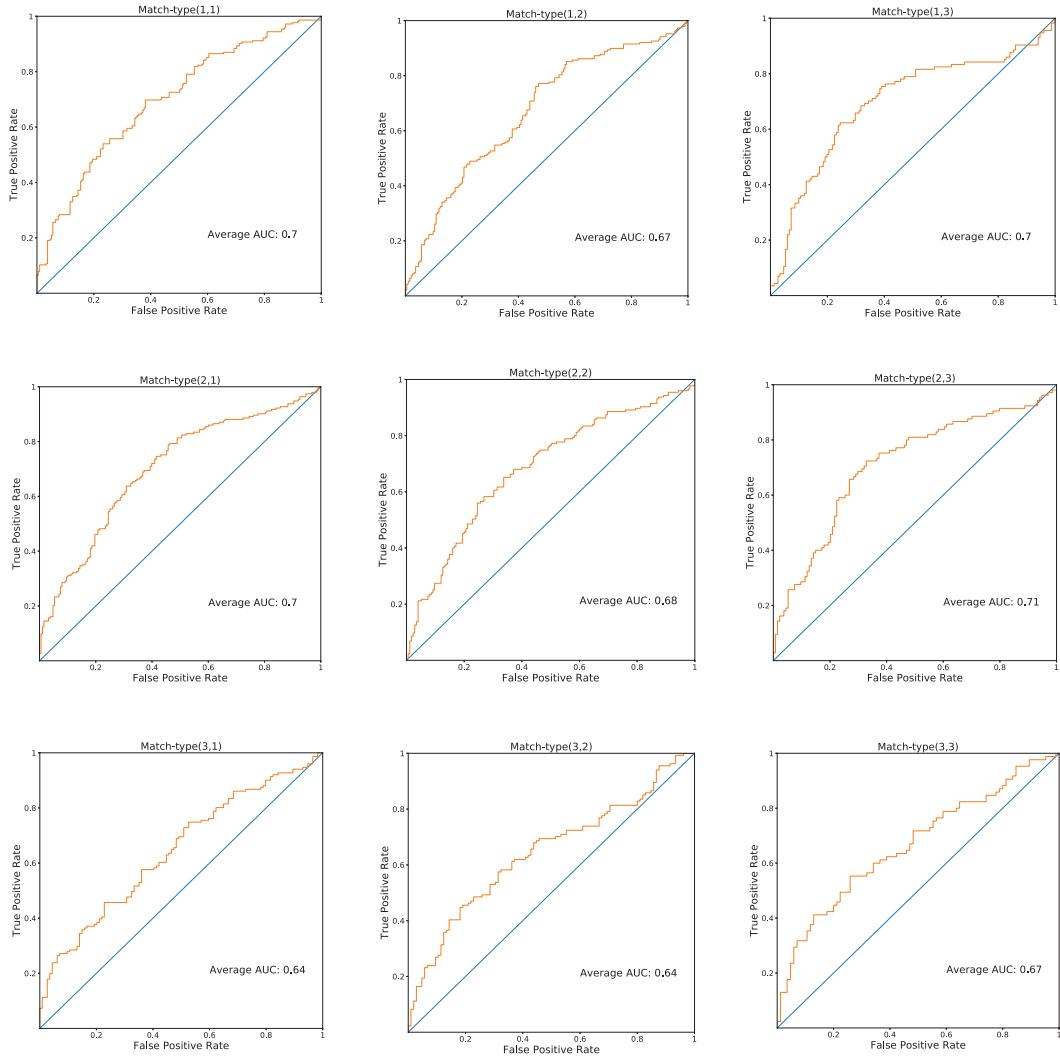


图 4.2: 整合模型用来区分测试集 T 和 TR 的ROC曲线和对应 AUC 值。

同时本论文还尝试用整合模型对 T 与 TR 做出区分（图4.2）。由于整合模型在区分 T 与随机产生的负测试的表现上不如 RBFN 模型，也可以推测整合模型在区分 T 与 TR 时不如 RBFN 模型。整合模型和 RBFN 模型在区分 T 与 TR 时的 AUC 平均差距为 7.1%。整合模型在区分 T 与随机产生的负样本上的表现（图3.3）也优于在区分 T 与 TR 时的表现，平均 AUC 差距为 12.8%。

更为有意思的是，这两个模型在区分 T 与 TR 以及 T 与随机负样本时的 AUC 差异似乎满足交换关系。为了更好的说明，先把各种情况下 AUC 的值列出来（表4.1）。

match-type	AUC(R,T,N)	AUC(R,T,TR)	AUC(I,T,N)	AUC(I,T,TR)
(1,1)	0.97	0.70	0.97	0.70
(1,2)	0.86	0.73	0.82	0.67
(1,3)	0.80	0.73	0.77	0.70
(2,1)	0.87	0.75	0.83	0.70
(2,2)	0.82	0.79	0.79	0.68
(2,3)	0.83	0.77	0.80	0.71
(3,1)	0.84	0.75	0.77	0.64
(3,2)	0.81	0.76	0.76	0.64
(3,3)	0.83	0.78	0.75	0.67

表 4.1: 各种情况下的AUC值。这些数值来自图 3.2, 3.3, 4.1, 4.2。R 表示 RBFN 模型, I 表示整合模型, T 表示正测试集, TR 表示由 T 中元素的抗原氨基酸序列和抗体氨基酸序列交换次序得到的测试集, N 表示随机生成的负样本。AUC(R,T,N) 表示用 RBFN 模型 (R) 区分 T 和 N 时的 AUC 值。

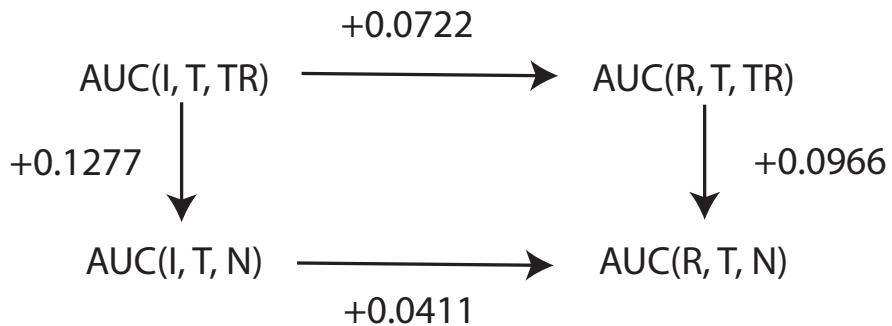


图 4.3: 各种情况下 AUC 的差值。0.0722 是 $AUC(I, T, R)$ 和 $AUC(R, T, R)$ 在各种 match-type 下差值的平均。0.1277, 0.0966, 0.0411 也做类似计算。

从图 4.3 可以看出

$$0.1277 + 0.0411 = 0.0722 + 0.0966$$

竟然是交换的(commutative)! 这个现象并不 trivial, 单单从数学的角度似乎得不到这样的等式, 同时通过比较在各种 match-type 下的数据也都不是严格成立。

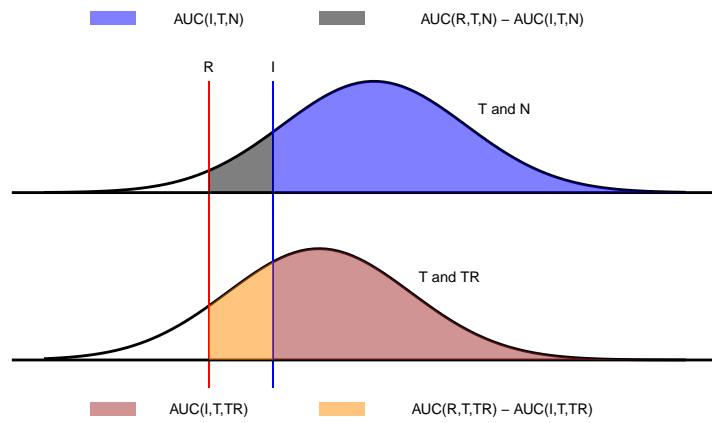


图 4.4: 对图 4.3 交换关系的解释

图 4.4 给出了交换关系的一个解释。这里接用心理学上阈限 (sensory threshold) 的概念，也就是两个刺激能够被感觉器官区分开的最小刺激强度的差。此处则用阈限来表示模型的好坏。那么，一个测试集中能够被区分开的样本的数目随着模型阈限的降低而增加。我们可以把测试集关于阈限的分布画出来，如图 4.4 中的两个曲线，分别表示测试集 T 和 TR 以及 T 和 N 沿阈限的分布；如图 4.4 中的两条直线分别表示 RBFN 模型(R)和整合模型(I)，它们在 x 轴上的截距表示这两个模型的阈限。显然整合模型的阈限大于RBFN的阈限，也就是RBFN模型要优秀。我们把两个分布曲线在直线右侧的面积简单理解成这个模型在对应测试集中的AUC，那么这样就可以解释图 4.3 的交换关系。而且，这两个模型在区分度比较差的测试集 (T 和 TR) 上的 AUC 值差距比较大，在区分度比较高的测试集上差距比较小。这里，所谓的区分度，可以理解为阈限在一定意义上的平均。如果上面的解释成立，那么当测试集的分布确定时候，我们就可以确定模型的差异。反过来，当模型的差异确定时，测试集之中个体的差异就有了一个衡量。

这里只是对交换性质做出的一个大胆解释，除了本文的例子之外还没有用其他例子来验证，但是本文强烈怀疑，这个解释对多数的模型和测试集都成立。

第 2 节 预测突变对抗原抗体复合物亲和力的影响

自然产生的抗体，往往在亲和力(affinity)上有一定的限制[51]。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[52]。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[53]，有不少的

亲和力(affinity)可以被提高到nM的级别[54]。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)[55],通过 potential-of-mean force(PMF)[56]，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[57]。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[58]，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[59]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^α 和 C^β 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[60]。

本论文从上面构建的RBFN模型出发，对突变的亲和力做出预测。一个基本的假设就是，如果突变使得抗原抗体相互作用的氨基酸朝着 Core 的方向变化，那么这个突变就可以增加抗原抗体的亲和力，否则就减弱抗原抗体的亲和力。那么，用 RBFN 模型进行亲和力变化预测的第一步就是找出相互作用的短的氨基酸序列。在现实的突变过程中，往往都是单个突变，然后逐次叠加，Sirin 等[59] 收集的数据中绝大多数为单位点突变（1101 个突变中有 645 个位单位点突变）。

2.1. 可预测氨基酸对的提取

突变氨基酸对的提取分为三个步骤：复合物的筛选，相互作用氨基酸对的提取，可预测氨基酸对的组装和筛选。

- (1) 复合物的筛选。在数据库 AB-Bind 中的突变，有的来自于在PDB中有三维结构的复合物，有的则来自于和PDB中一些复合物相似的复合物。由于在相互作用的氨基酸对的提取过程中，需要抗原抗体复合物，我们则之选取哪些来自PDB中复合物的突变。一共有以下 17 个。

1bj1, icz8, 1dqj, 1mhp, 1mlc, 1vfb, 1yy9, 2jel

2ny7, 2nyy, 2nz9, 3bdy, 3be1, 3bn9, 3hfm, 3ngb, 1dvf

- (2) 相互作用突变氨基酸对的提取。本论文在前面抽取 Core 的时候，选取在截断距离是 4\AA 的时候，CN 数值最大的那个氨基酸序列对。这里，由于我们的目的不是抽取 Core，而是判断突变后，氨基酸对是不是更趋向 Core，所以我们的截断距离选用了移动的截断距离。这里，截断距离的变化范围从 3\AA 变化到 8\AA ，且步长是 0.25。对于每一个突变位置，通过如图 4.5 的过程找出相互作用的氨基酸对。

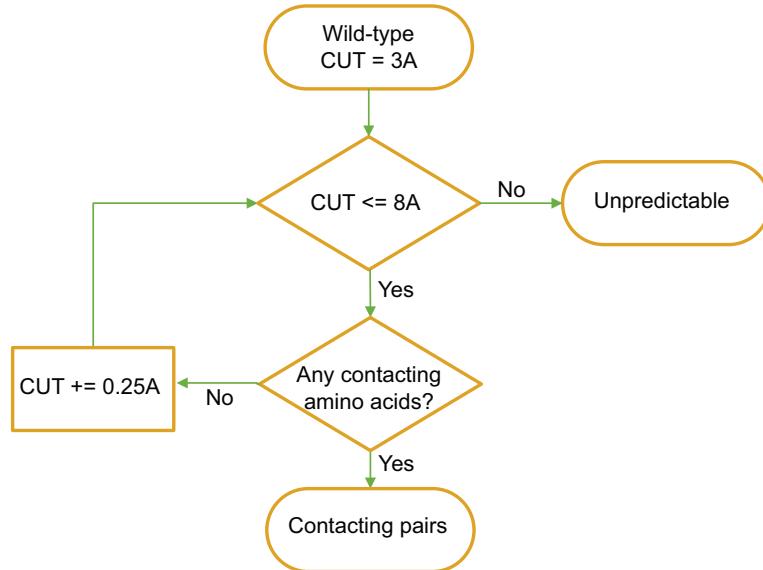


图 4.5: 相互作用突变氨基酸对的提取。

- (3) 可预测氨基酸对的组装和筛选。本论文对上一步得到的相互作用的氨基酸对进行了六种类型的组装和筛选。

- (a) one_WithinRange_False : 对于同一个突变复合物中所有突变氨基酸对，找出 CN 最大的一个作为预测亲和力的依据。这里所谓一个突变复合物是指有一个或者多个突变并且突变后抗原抗体复合物的亲和力通过实验测得。称一个突变复合物的野生型为野生复合物。
- (b) one_WithinRange_True 对于同一个突变复合物中所有突变氨基酸对，找出 CN 最大的一个，同时，这个 CN 最大的氨基酸对中的抗体氨基酸必须位于 CDR 区域内。这里的 CDR 区域和本文提取 Core 时定义的 CDR 区域一致。选取规则可以概括为图 4.6。

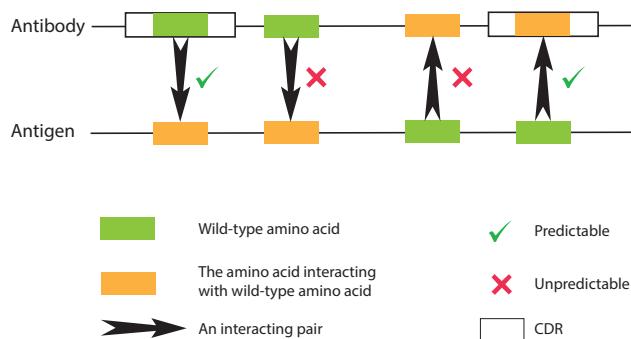


图 4.6: 把突变氨基酸对的抗体氨基酸限定在 CDR 上。

- (c) multiple_WithinRange_False : 同一个突变复合物中所有突变氨基酸对都作为判断依据。
- (d) multiple_WithinRange_True : 对同一个突变复合物中所有突变氨基酸对按照图 4.6 中的规则进行筛选。如果这个突变复合物中有一个突变氨基酸对不满足要求，那么这个突变复合物中的所有氨基酸对都不可预测。
- (e) flanked_WithinRange_False : 当 multiple_WithinRange_False 的氨基酸对选定后，每个氨基酸连同它前面和后面的氨基酸一起考虑。例如，([Ab1], [Ag1]) 是 multiple_WithinRange_False 中选定的可预测的氨基酸对，那么 ([Ab0, Ab1, Ab2], [Ag0, Ag1, Ag2]) 就是 flanked_WithinRange_False 中选定的可以预测的氨基酸对。
- (f) flanked_WithinRange_True : 当 multiple_WithinRange_True 的氨基酸对选定后，每个氨基酸连同它前面和后面的氨基酸一起考虑。

2.2. 亲和力变化的预测

当可以预测的氨基酸对提取出来之后，属于同一个突变复合物的氨基酸对放在一起。

假设 $(Mu_1, Ag_1), (Mu_2, Ag_2), (Ab_1, Mu_3)$ 为一个突变复合物中的可预测氨基酸对， $(Wt_1, Ag_1), (Wt_2, Ag_2), (Ab_1, Wt_3)$ 为对应的野生型复合物中的可预测氨基酸对。这里可预测氨基酸对可以是六种类型中的任意一种 (one_WithinRange_True , one_WithinRange_False , multiple_WithinRange_True , multiple_WithinRange_False , flanked_WithinRange_True , flanked_WithinRange_False) 。那么，突变后亲和力的变化由下面的公式给出判断：

$$\begin{aligned} RBFN_{Mu} &= \frac{1}{3}[RBFN(Mu_1, Ag_1) + RBFN(Mu_2, Ag_2) + RBFN((Ab_1, Mu_3))]3 \\ RBFN_{Wt} &= \frac{1}{3}[RBFN(Wt_1, Ag_1) + RBFN(Wt_2, Ag_2) + RBFN((Ab_1, Wt_3))]3 \\ \Delta &= RBFN_{Mu} - RBFN_{Wt} \end{aligned}$$

如果 $\Delta > 0$ 则说明亲和力增加；如果 $\Delta < 0$ 则说明亲和力降低。

Mode	$\Delta\Delta G < 0.5$	$\Delta\Delta G > 0$	$\Delta\Delta G > 0.5$	$\Delta\Delta G > 1$	n
one_WithinRange_True	0.531	0.605	0.620	0.710	467
one_WithinRange_False	0.526	0.603	0.619	0.713	497
multiple_WithinRange_True	0.525	0.599	0.606	0.693	463
multiple_WithinRange_False	0.524	0.598	0.605	0.695	497
flanked_WithinRange_True	0.544	0.506	0.427	0.458	463
flanked_WithinRange_False	0.529	0.489	0.408	0.429	497

表 4.2: $\Delta\Delta G$ 位于不同范围的突变复合物的亲和力的预测效果。其中 $\Delta\Delta G$ 为吉布斯自由能的变化，单位是 KJ/mol 。 n 表示预测的突变复合物的数目。预测效果的好坏用 AUC 来衡量。

从表格4.2 可以看出，one_WithinRange_True 和 one_WithinRange_False 为预测效果最好的两个，而且这两者，除了 $\Delta\Delta G > 1$ 的情况外，one_WithinRange_True 类型的预测效果稍优，但是优势不明显。其中一个原因就是多数情况下，相互作用的氨基酸对都满足图 4.6 的要求。由于本文的模型是建立在 CDR 上的氨基酸和抗体氨基酸之间相互作用的基础上，为了保持一致，选择类型 one_WithinRange_True 对突变复合物亲和力的变化进行预测。

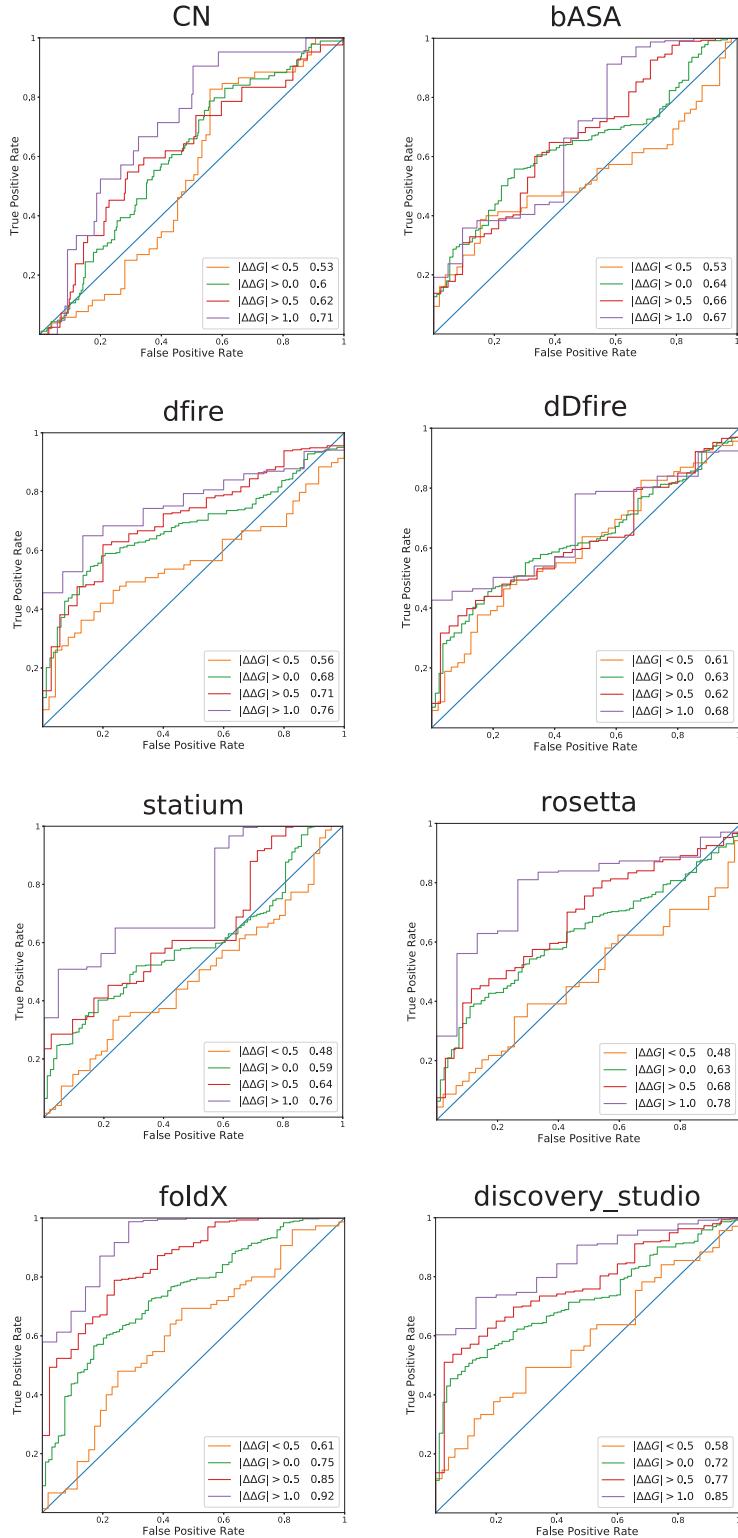


图 4.7: 各种方法的预测效果。CN: 指的是本文的方法, 是 Contacting Number 首字母的缩写。其他的方法请参考文献 [59]。突变体复合物的选择是在 one_WithinRange_True 的类型下的突变体复合物。 $\Delta\Delta G$ 的单位是 KJ/mol。

图 4.7 给出了不同方法的预测效果。从图中可以看出，本文的方法 CN 和除 foldX 以及 Discovery Studio 外的其他五种方法在不同的 $\Delta\Delta G$ 的子集上互有优劣。为了准确的衡量本文的方法与其他方法相较之下的优劣，本文用 Bootstrap 的方法构建了 AUC 的置信区间，如表 4.3。从中可以看出，Discovery Studio 在 $\Delta\Delta G > 0$, FoldX 在 $\Delta\Delta G > 0, \Delta\Delta G > 0.5, \Delta\Delta G > 1$ 都优于 CN，所有其他情况和方法，在95%置信区间意义下都不比CN有明显优势。综合以上结果，bASA, dfir, dDfir, Rosetta, STATIUM 和 CN 在 one_WithinRange_True 得到的测试集中，旗鼓相当。

	$\Delta\Delta G < 0.5$	$\Delta\Delta G > 0$	$\Delta\Delta G > 0.5$	$\Delta\Delta G > 1$
CN	(0.43, 0.63)	(0.54, 0.66)	(0.53, 0.71)	(0.60, 0.80)
bASA	(0.44, 0.64)	(0.58, 0.69)	(0.57, 0.74)	(0.54, 0.80)
dfire	(0.45, 0.67)	(0.62, 0.73)	(0.63, 0.79)	(0.67, 0.84)
dDfir	(0.50, 0.71)	(0.57, 0.68)	(0.54, 0.70)	(0.57, 0.78)
Rosetta	(0.37, 0.68)	(0.57, 0.68)	(0.59, 0.76)	(0.67, 0.87)
STATIUM	(0.42, 0.63)	(0.57, 0.68)	(0.58, 0.75)	(0.68, 0.87)
D Studio	(0.48, 0.69)	(0.67, 0.77)	(0.70, 0.83)	(0.77, 0.92)
FoldX	(0.51, 0.71)	(0.69, 0.8)	(0.79, 0.91)	(0.86, 0.98)

表 4.3: Bootstrap 方法构建各种不同方法 AUC 95% 的置信区间。Bootstrap 的取样次数为 10,000 。 D Studio 为 Discovery Studio 。

第五章 讨论

本论文的核心就是通过对抗原抗体相互作用的 Core, 也就是抗原抗体相互作用的关键的氨基酸的研究, 建立了抗原和抗体相互作用的一个数量关系。这是第一次。在这之前的研究都是零散的, 不成体系的, 很难推广的。其中最主要的原因是抗原抗体的相互作用的氨基酸是一个空间的关系, 而不是线性的关系, 而本文从 Core 的角度成功的绕过了这个障碍。本文的模型在区分 Core 中的抗原氨基酸和抗体氨基酸, 以及预测突变对抗原抗体复合物亲和力的影响上, 取得一定的成功。

用 Core 的概念来研究抗原抗体的相互作用, 据本论文所了解的知识范围来讲, 也是第一次。Core 的概念是建立在另外一个重要的概念之上: CN(contact number)。当我们按照 4\AA 的距离进行截断, 得到相互作用的氨基酸后, 本论文用 CN 的大小来衡量相互作用的强弱。这个是首次, 区别于以往接触面积的度量方法[80, 81]。按照 CN 方法来定以关键的相互作用的氨基酸在直观上是合情合理的, 最后建立的模型效果也对 CN 的有效性做了间接的证明。其实, 也可以通过分析实验数据来对 CN 的有效性做最直接的判定。

本论文模型成立的另外一个重要假设是 BLOSUM62 可以综合的反应氨基酸之间性质的差异。有很多文章通过单独研究某一种类型的相互作用, 比如氢键, 疏水作用, 盐桥等, 但是这些单独的相互作用很难对氨基酸的相互作用做一个全面的描述, 同时也很难说一个全面的描述应该包含哪些类型的相互作用, 还有就是这些单独的相互作用很难量化。本文认为, BLOSUM62 对氨基酸之间的差异做了一个全面综合的描述, 因为 BLOSUM62 是在分析了大量同源氨基酸序列之间的替换关系后得到的。就好比一辆汽车, 可以通过分析燃料的能量值、燃料的燃烧效率、摩擦力、汽车本身的重量、汽车的外形等得到最终汽车的拉力, 也可以直接让汽车跑一跑, 直接测得汽车得拉力。

在建立 RBFN 模型得时候, 有一个重要得环节, 就是中心得选取。本论文采用了一个计算上更有效得由文献 [71] 给出的方法。这个方法最终的确定, 也经历了很多的探索。第一个要问的问题, 有没有更好的方法? 结合对径向基函数网络在生物上的应用的思考, 本论文给出了下面初步的探索。

用数学来研究生物问题, 第一件要做的事情就是把生物对象之间的关系用数字来描述, 这是数学模型的开始。距离, 作为描述相互关系的一个指标, 在生物问题上比较容易获得。比如说蛋白质之间的距离, 可以通过它们所在的蛋白质作用网络来描述, 也可以通过序列的相似性, 由BLOSUM62来刻画。再比如说DNA序列之间的距离可以通过各种模型下核酸之间的替换概率来计算[73, 74, 75, 76]。再比如, 生物网络各个节

点之间的关系也可以通过距离来衡量。当有了距离之后，径向基函数网络就显得自然而然，因为它的输入值就是距离。在构建径向基函数网络的时候，一个重要的步骤就是中心(center)的选取。当中心选定之后，对每个样本的描述都由它和这些中心之间的关系决定。所以，选择好的中心，可以有效的压缩和描述数据。在生物学上，这些中心应该是具有重要生物意义的对象。比如，在研究癌症和基因之间的关系时，好的中心可能是对癌症产生重要影响的基因。Alexandridis 等给出了一个高效的选取 CRBFN 中心的方法[71]。基本思想就是对样本进行聚类，然后从每个聚类中随机抽取中心。这个方法受到随机因素的影响，特别是抽取的中心数目比较少时，这种随机抽取的方法更是容易漏掉真正有重要意义的中心。所以，如果能在损失函数的指导下抽取中心，情况就会好很多。

假设损失函数是 $L(C, W)$ ，其中 $C = (c_1, c_2, \dots, c_n)$ 是 n 个中心， $W = (w_1, w_2, \dots, w_n)$ 是对应于这 n 个中心的系数。那么中心 c_i 的重要性可以由 $|\frac{\partial L}{\partial w_i}|$ 的大小来定义， $|\frac{\partial L}{\partial w_i}|$ 越大， c_i 在模型中越重要。但是，优化模型时， $|\frac{\partial L}{\partial w_i}|$ 趋于 0，这样就不容易对不同中心的重要性做有效的区分。为了解决这个问题，可以对损失函数加上一个很小的正则项。

$$L_\delta = L(W, C) + \frac{\delta}{2} \sum w_i^2 \quad (5.1)$$

然后按照新的损失函数 L_δ 来训练模型，则有

$$\frac{\partial L_\delta}{\partial w_i} = \frac{\partial L(W, C)}{\partial w_i} + \delta w_i = 0 \quad (5.2)$$

$$\delta |w_i| = \left| \frac{\partial L_\delta}{\partial w_i} \right| \quad (5.3)$$

其中 $\delta > 0$ ，是一个常数。公式 (5.1), (5.2), (5.3) 就是说：按 $|\frac{\partial L}{\partial w_i}|$ 排序和按 $|w_i|$ 排序的结果一样，也就是：

$$\left| \frac{\partial L}{\partial w_i} \right| > \left| \frac{\partial L}{\partial w_j} \right| \iff |w_i| > |w_j| \quad (5.4)$$

依据公式 (5.4)，有算法 1，称这个方法为 LGM(Loss Guided Method)。

本论文采用了 Hamming, Occurrence Frequency[78], Inverse Occurrence[78] 和 Eskin[79], Frequency 四种距离，Gaussian, Inverse Multi-Quadric, Markov 和 Thin Plate Spline 四种径向基函数，在 Breast Cancer 数据上 [77]，对 LGM 和文献 [71] 的方法（这里称之为 Random Cluster）进行比较。比较的标准采用 MCC(Matthews correlation coefficient)，结果如图 5.1, 5.2, 5.3, 5.4。从这些图中可以看出，当中心的数目比较多时，LGM 和 Random Cluster 表现趋于一致，但是当中心的数目被压缩到很少，LGM 的表现要明显好于 Random Cluster。

这里只给出了在 Breast Cancer 数据上的比较结果，在本论文中 Core 的数据上，也有同样的结论，也就是当中心的数目比较少时，LGM 有明显优势，当中心数目比较

Algorithm 1 Center Selection Method

```

Initial  $W$                                  $\triangleright W$  is a vector of coefficients
Initial  $C$                                  $\triangleright C$  begins with all samples.
Set the target number of centers  $N$ .
while  $\text{len}(C) < N$  do                 $\triangleright \text{len}(C)$  is the length of  $C$ .
    Train the model
    for  $w_i \in W$  do
        if  $|w_i| = \min\{|w| : w \in W\}$  then
            Del  $w_i$  from  $W$ 
            Del  $c_i$  from  $C$ 
        end if
    end for
end while

```

多的时候，二者表现趋同。而且可以预料，这个结论是普遍成立的，因为 LGM 是有目的的去选择中心。但是本论文的目的是构建好的模型，而不是选则最少的中心，所以采用了文献 [71] 的方法，因为这种方法在计算上更高效。这也是 LGM 的一个重大缺点，特别是当数据集比较大的时候，LGM 显得笨拙。但是，这个问题可以采取折中的方法来解决，可以先采用文献 [71] 的方法筛选出一部分中心，然后再利用 LGM 接着筛选。

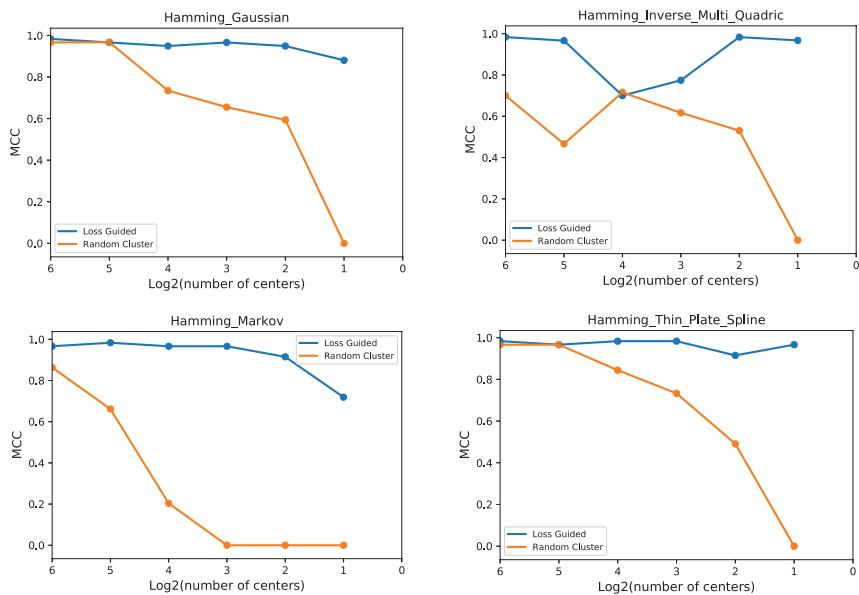


图 5.1: Hamming distance 的定义可以参看文献[71]。Loss Guided 表示 LGM 方法，Random Cluster 表示文献 [71] 给出的方法。

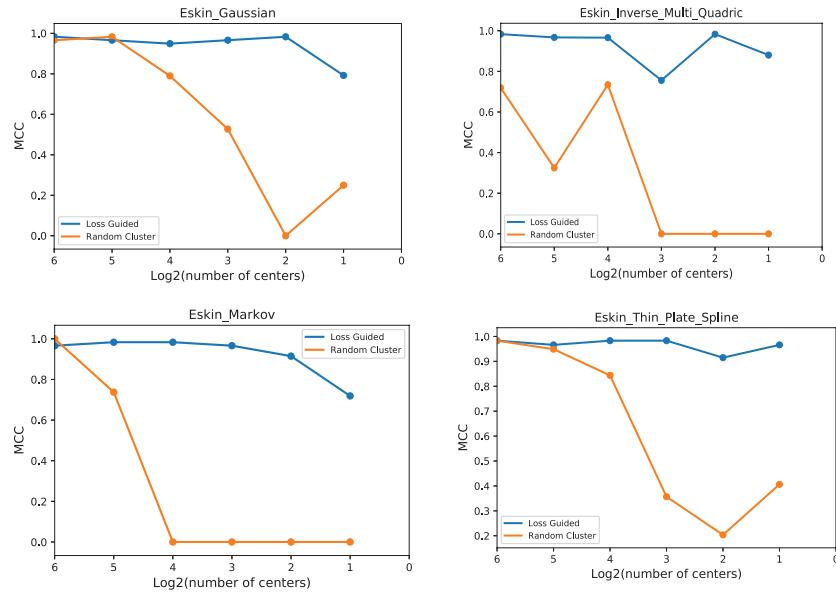


图 5.2: Eskin distance 的定义可以参看文献[79]。Loss Guided 表示 LGM 方法， Random Cluster 表示文献 [71] 给出的方法。

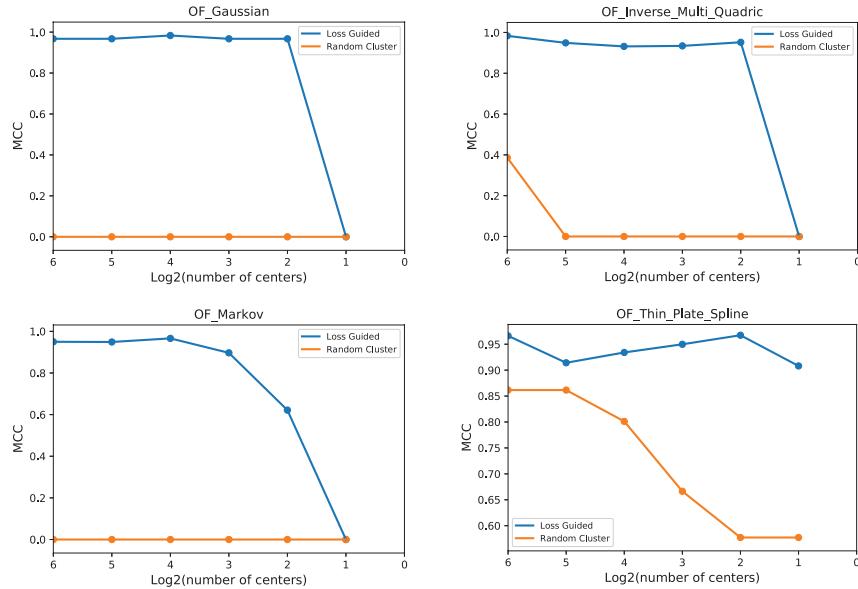


图 5.3: OF 表示 Occurrence Frequency , 定义可以参看文献[78]。Loss Guided 表示 LGM 方法， Random Cluster 表示文献 [71] 给出的方法。

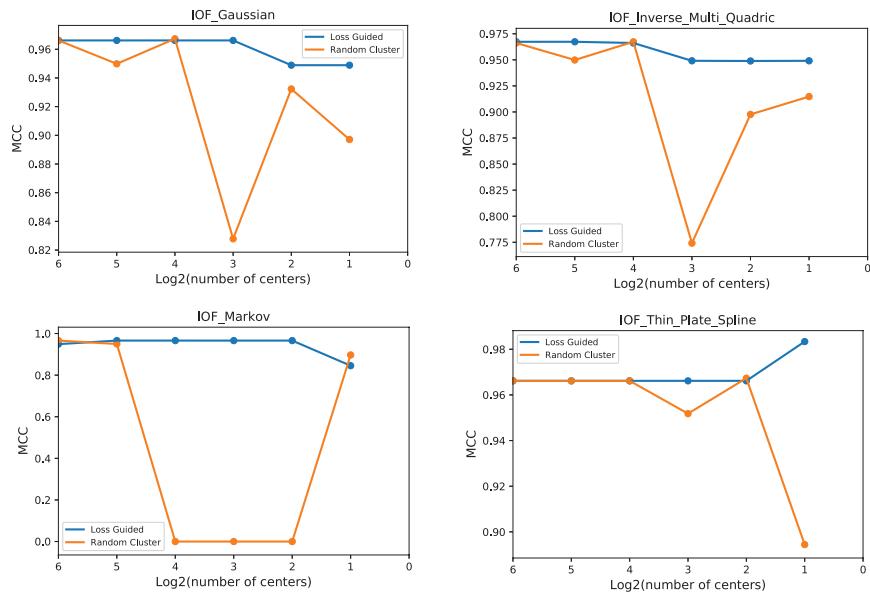


图 5.4: IOF 表示 Inverse Occurrence Frequency , 定义可以参看文献[78]。Loss Guided 表示 LGM 方法, Random Cluster 表示文献 [71] 给出的方法。

在预测突变对抗原抗体亲和力的影响时, 不弱于其他的许多方法(STATIUM, Rosetta, dfire, dDfire, rASA), 但是也没有本质的提升, 和其他的许多方法一样表现平平, 不能提供精准的预测, 只能在某些时候, 为现实的应用提供少许帮助。

最后要说的是, 本论文在创新的解决了抗原抗体相互作用领域的一些重要问题的同时, 也提出了很多问题。CN 在判断相互作用强弱上的有效性可不可以通过实验测得的数据进行验证? 虽然预计本文的模型被用来预测抗原表位会提高预测的准确性, 能不能设计一种方法试试? Core 反映了短的氨基酸序列之间的相互作用, 但是归根结底, 抗原抗体的相互作用还是空间上的, 能不能把 Core 的概念推广到空间上, 建立更符合实际的模型? 这些都是具有重要意义的问题, 也都是可以解决的问题, 并不是遥不可及。

本论文有关的数据和代码, 请参看

https://github.com/Leochuanxing/Paritope_Epitope

致谢

在本论文相关的研究过程中，得到了朱山风老师和孙丰珠老师的大量帮助。在此表达诚挚的感谢。还要感谢 MIT 的 Keating 教授，她无私的提供了她们组的实验数据。感谢吴宗敏老师在数学上的指导。还要感谢过去并没有直接参与到本论文的老师和朋友们，虽然他们没有直接参与论文的写作，但是他们的存在使我更像一个社会性的动物，让我的存在有一定的意义。

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] Dorfman T, Moore MJ, Guth AC, Choe H, Farzan M. A tyrosine-sulfated peptide derived from the heavy-chain CDR3 region of an HIV-1-neutralizing antibody binds gp120 and inhibits HIV-1 infection. *Journal of Biological Chemistry*. (2006), 281(39):28529-35.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.

- [16] 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.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] Ekiert, Damian C., et al. Antibody recognition of a highly conserved influenza virus epitope. *Science*, (2009), 324(5924): 246-251.
- [24] , S Ferdous, S Kelm, TS Baker, J Shi, ACR Martin, B-cell epitopes: Discontinuity and conformational analysis *Molecular immunology*, (2019), 114:643-650.
- [25] G. E. Morris, Epitope mapping, *Methods in Molecular Biology*, (2005), 295:255 – 268.
- [26] Rubinstein, Nimrod D., et al. Computational characterization of B-cell epitopes. *Molecular immunology*, (2008), 45(12):3477-3489.
- [27] Hopp, T. P., Woods, K. R. Prediction of protein antigenic determinants from amino acid sequences. *PNAS*, (1981), 78(6), 3824-3828.
- [28] Rubinstein, N. D., Mayrose, I., Pupko, T. A machine-learning approach for predicting B-cell epitopes. *Molecular immunology*, (2009), 46(5), 840-847.
- [29] 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.
- [30] Kulkarni-Kale, U., Bhosle, S., Kolaskar, A. S.,CEP: a conformational epitope prediction server. *Nucleic acids research*, (2005), 33(suppl_2), W168-W171.
- [31] 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.
- [32] 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.
- [33] Ponomarenko, J., Bui, H. H., Li, W., Fusseder, 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.

- [34] 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.
- [35] 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.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] Sela-Culang, I., Ashkenazi, S., Peters, B., Ofran, Y. PEASE: predicting B-cell epitopes utilizing antibody sequence. *Bioinformatics*, (2015), 31(8), 1313-1315.
- [40] 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.
- [41] 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.
- [42] 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.
- [43] 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.
- [44] Coenen F, Goulbourne G, Leng P, Tree Structures for Mining Association Rules. *Data Min. Knowl. Discov.*, (2004), 8:25-51
- [45] 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.
- [46] Clackson T , Wells J . A hot spot of binding energy in a hormone-receptor interface. *Science*, (1995), 267(5196):383-386.
- [47] 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.
- [48] 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.
- [49] 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.
- [50] 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.
- [51] Foote, J., Eisen, H.N. Kinetic and affinity limits on antibodies produced during immune responses. *Proc. Natl. Acad. Sci*, (1995), 92, 1254-1256.

- [52] 5 Presta LG, Selection, design, and engineering of therapeutic antibodies. *J Allergy Clin Immunol*, (2005), 116(4):731-736
- [53] 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.
- [54] 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.
- [55] Chipot C, Pohorille A. Calculating Free Energy Differences Using Perturbation Theory. Free Energy Calculations *Springer Series in CHEMICAL PHYSICS: Springer*, (2007). p. 33 - 75.
- [56] Roux B. The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, (1995);91(1 - 3):275 - 82
- [57] 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.
- [58] 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.
- [59] Sirin, Sarah, et al. AB-Bind: antibody binding mutational database for computational affinity predictions. *Protein Science*, (2016), 25(2): 393-409.
- [60] 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.
- [61] 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.
- [62] Allazikani B, Lesk A M, Chothia C. Standard conformations for the canonical structures of immunoglobulins. *Journal of Molecular Biology*, (1997), 273(4):927-948.
- [63] Fellouse, F. A., Barthelemy, P. A., Kelley, R. F., and Sidhu, S. S. Tyrosine plays a dominant functional role in the paratope of a synthetic antibody derived from a four amino acid code. *Journal of molecular biology*, (2006), 357, 100 - 114.
- [64] Tsumoto, K., Ogasahara, K., Ueda, Y., Watanabe, K., Yutani, K., and Kumagai, I. Role of Tyr residues in the contact region of antilysozyme monoclonal antibody HyHEL10 for antigen binding. *Journal of Biological Chemistry*, (1995), 270, 18551 - 18557.
- [65] Sara Birtalan, Yingnan Zhang, Frederic A. Fellouse, The Intrinsic Contributions of Tyrosine, Serine, Glycine and Arginine to the Affinity and Specificity of Antibodies. *Journal of Molecular Biology*, (2008), 377(5):0-1528.
- [66] Fellouse, F. A . Synthetic antibodies from a four-amino-acid code: A dominant role for tyrosine in antigen recognition. *Proceedings of the National Academy of Sciences*, (2004), 101(34):12467-12472.
- [67] Fellouse, Frederic A. , et al. Tyrosine Plays a Dominant Functional Role in the Paratope of a Synthetic Antibody Derived from a Four Amino Acid Code. *Journal of Molecular Biology*, (2006), 357(1):0-114.
- [68] Young, L., Jernigan, R. L., and Covell, D. G. A role for surface hydrophobicity in protein-protein recognition. *Protein Science*, (1994), 3, 717 - 729.
- [69] Bhat, T. N., Bentley, G. A., Boulot, G., Greene, M. I., Tello, D. W. D. A., Dall' Acqua, W., Souchon, H., Schwarz, F. P., Mariuzza, R. A. and Poljak, R. J. Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proceedings of the National Academy of Sciences*, (1994), 91, 1089 - 1093.

- [70] Dash P , Fiore-Gartland A J , Hertz T , et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, (2017), 547(7661):89-93.
- [71] Alexandridis, Alex and Chondrodima, Eva and Giannopoulos, Nikolaos and Sarimveis, Haralambos. A fast and efficient method for training categorical radial basis function networks. *IEEE transactions on neural networks and learning systems*, (2016), 28(11):2831-2836.
- [72] Jorge Nocedal Stephen J. Wright, Numerical Optimization, Springer Series in Operations Research and Financial Engineering, 2006.
- [73] Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, (1980), 16(2):111-20.
- [74] Hasegawa M, Kishino H, Yano TA. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution*, (1985), 22(2):160-74.
- [75] Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences*, (1986), 17(2):57-86.
- [76] Yang Z. Estimating the pattern of nucleotide substitution. *Journal of molecular evolution*, (1994), 39(1):105-11.
- [77] A. Asuncion and D. J. Newman, UCI Machine Learning Repository. Irvine, CA, USA: Univ. California Irvine, 2007.
- [78] K. S. Jones, A statistical interpretation of term specificity and its application in retrieval, in Document Retrieval Systems, W. Peter, Ed. London, U.K.: Taylor Graham Publishing, 1988, pp. 132 - -142.
- [79] S. Boriah, V. Chandola, and V. Kumar, Similarity measures for categorical data: A comparative evaluation, in *Proc. 8th SIAM Int. Conf. Data Mining*, 2008, pp. 243 - 254.
- [80] Conte LL, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. *Journal of molecular biology*, (1999), 285(5):2177-98.
- [81] Levy, E.D., A simple definition of structural regions in proteins and its use in analyzing interface evolution. *Journal of molecular biology*, (2010), 403(4), pp.660-670.
- [82] T. Hastie et al., The Element of Statistical Learning, Springer Series in Statistics, Springer-Verlag, 2001.
- [83] S. Chen, Mach configuration in pseudo-stationary compressible flow, *J. Amer. Math. Soc.*, 21(2008), no. 1, pp. 63–100.
- [84] 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.
- [85] 陈纪修, 潘崇华, 金路, 数学分析, 高等教育出版社, 1999.
- [86] 苏步青, 数学教育与应用数学问题, 数学通报, 1988, (2): 1-2.
- [87] 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.