特异识别 PAM 序列的 CRISPR-Cas9 突变体的计算设计

完成人
段致远

指导老师

黄强　教授

目　录

摘　要

SpCas9 蛋白是一种源于原核生物的核酸内切酶，可以在 RNA 的引导下特异性切割有互补序列的双链 DNA。虽然 SpCas9 已被广泛应用于基因编辑，但其应用范围受限于对目标DNA旁的一段PAM序列(5’-NGG-3’)的识别。通过对SpCas9进行改造使其识别不同的 PAM 可以丰富其在基因编辑的应用。基于野生型 SpCas9 的结构，我们希望使用计算结构生物学的方法设计识别不同 PAM 序列的 SpCas9 突变体。根据对 SpCas9 晶体结构的研究，我们自定义了筛选突变体的评分函数并通过实验已经发现的突变体验证了该评分函数的有效性。在模拟中，我们发现了一系列能够提升识别 5’-NGC-3’、5’-NGT-3’和 5’-NCC-3’特异性的突变。我们通过动力学模拟分析了识别 5’-NGC-3’特异性突出 PVHLE 突变体，但结果表明该识别并不稳定。本研究的结果可以为进一步设计特异识别 PAM 的 SpCas9 突变体提供参考。

关键词：

SpCas9，PAM，计算结构生物学，蛋白质-DNA 设计

Abstract

The RNA-guided endonuclease Cas9, generated from prokaryote, cleaves double-stranded DNA with complementary sequence. Although SpCas9 has been widely used for genome editing, the use is constrained by the need of a specific PAM sequence on target DNA. Engineering Cas9 with altered PAM specificity may drop this limitation and expand genome editing application. Based on crystal structure of SpCas9, we tried to engineer SpCas9 with designate PAM by computational redesign. We adopted hydrogen-bonding preferred score-function to select variants and tested the credibility through control group which contained experimental verified mutants. Some amino acid mutations were selected for targeting 5'-NGC-3', 5'-NGT-3' and 5'-NCC-3' PAMs. The stability of PVHLE mutant, targeting 5'-NGC-3', was verified by dynamic simulation but the outcome was negative. Our work provides reference for further redesign of SpCas9 targeting novel PAM.

Keywords,
SpCas9, PAM, computational structure biology, protein-DNA interface design

一、前 言

CRISPR/Cas 系统是在原核生物中发现的一种抵御外源 DNA 入侵的免疫机制。该系统的一些 Cas 蛋白具有特异切割 DNA 或 RNA 的功能，其中 II 型 CRISPR-Cas 系统中的 Cas9 蛋白质，由于能在单链 RNA 的指导下特异切割 DNA 双链而被广泛应用于基因编辑[1, 2]。将 Cas9 基因和针对靶基因设计的编码 gRNA 的基因一起转化受体细胞，Cas9 表达后通过 gRNA 定向并切割靶序列，经过 DNA 非同源末端的修复，即可实现在特定靶位点对基因进行编辑[3, 4]。

Cas9/RNA 复合体靶向目标 DNA 序列不仅依赖 gRAN 与 DNA 靶序列约 20 个互补碱基的配对，而且也依赖 Cas9 蛋白对 DNA 互补序列旁一段特异性序列的识别，称为 PAM（proto-spacer adjacent motif）序列[5]。目前应用最广泛的是 SpCas9（*Streptococcus pyogenes* Cas9），该蛋白所识别的 PAM 序列为 5'-NGG-3'（图 1.1）。由于 PAM 序列的识别对于 Cas9 蛋白质对目标 DNA 序列的切割是必须的，因此在进行基因编辑时不仅需要设计与目标基因序列互补的 gRNA，还要考虑目标基因序列是否有相邻的 PAM 序列。这种限制使得对较短的基因编辑变得非常困难，如 microRNAs、短的开放阅读框、转录因子结合位点等[6, 7]。因此丰富 Cas9 的种类以识别不同的 PAM 序列可以扩大基因编辑的应用范围。

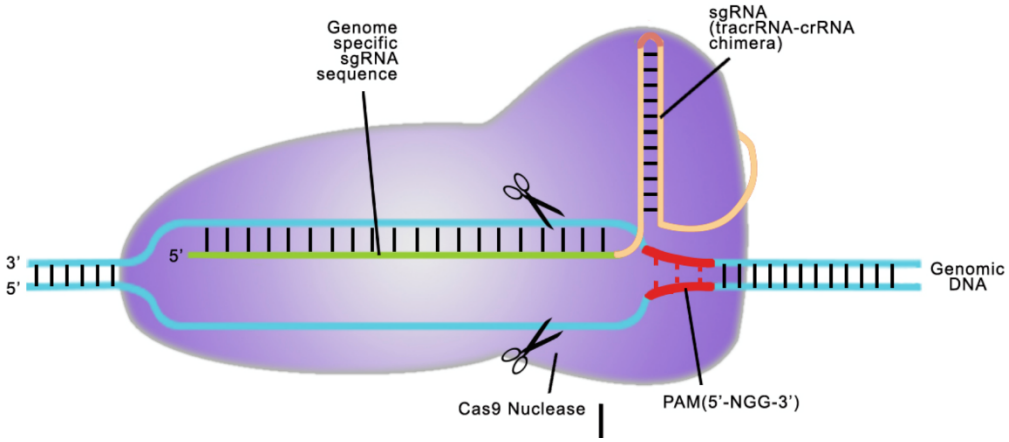

图 1.1. SpCas9/RNA/DNA 复合体示意图。SpCas9 在 RNA 的指导下特异性靶向互补的 DNA 序列，在互补序列旁有 PAM 序列。（图片来源：esternmorningherald.com/crispr-cas9/1240836）

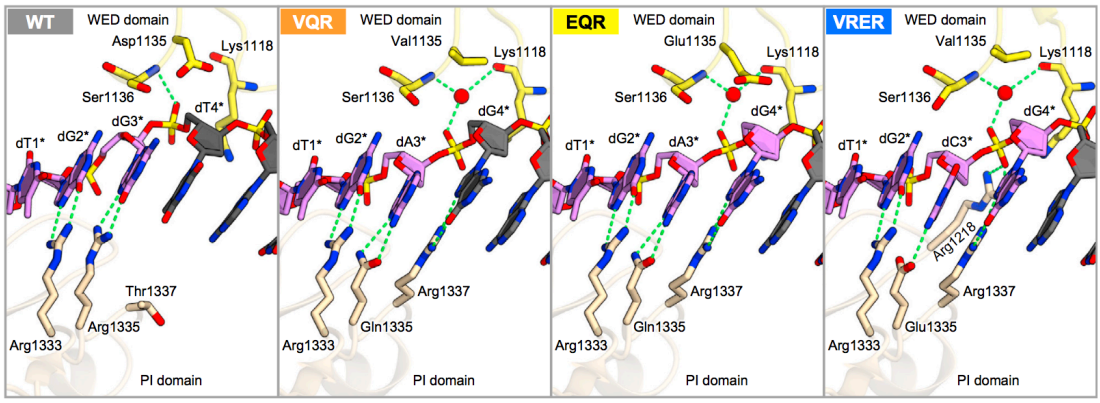

图 1.2. SpCas9 突变体。野生型 SpCas9 识别 5’-NGG-3’的 PAM（左一）；VQR 突变体识别 5’-NGAG-3’的 PAM（左二）；EQR 突变体识别 5’-NGAG-3’的 PAM（右二）；VRER 突变体识别 5’-NGAG-3’的 PAM（右一）。（Hirano S., 2016）

目前，已有研究人员通过随机突变实验成功筛选了一系列识别不同 PAM 的 SpCas9 突变体，并对其中识别特异性稳定的突变体解析了晶体结构，这些突变体是识别 5’-NGA(G)-3’的 VQR（D1135V / R1335Q / T1337R）突变体、识别 5’-NGAG-3’的 EQR（D1135E / R1335Q / T1337R）突变体和识别 5’-NGCG-3’的 VRER（D1135V / G1218R / R1335E / T1337R）突变体（图 1.2）[8]。这些突变体除了直接识别氨基酸的突变，还有间接识别氨基酸的突变。根据对这些晶体结构的研究，间接识别氨基酸突变对蛋白质和 DNA 主链的空间位置有显著的影响，如 1135 位的 Val 突变。该突变通过影响相互识别的氨基酸与核酸间的距离，改变了氨基酸与核酸的相互作用的强弱及 SpCas9 的识别特异性[9]。

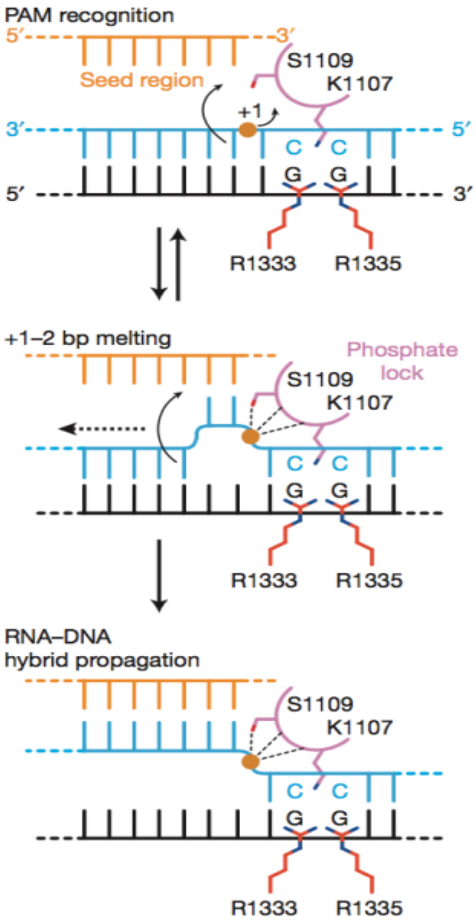

图 1.3. 在 SpCas9/RNA/DNA 复合体形成过程中 Cas9 识别 PAM 的分子机制模型。(Anders C, 2014)

在对 SpCas9/RNA/DNA 复合体晶体结构的研究中，人们发现 Cas9 与 RNA、DNA 形成复合体并发挥切割 DNA 的功能涉及一系列结构变化。其中 Cas9 蛋白识别 PAM 序列并引发 DNA

双链解旋是复合体形成的关键步骤。由晶体结构得到的理论模型提出：DNA 的解旋过程中 1333Arg 与 1335Arg 稳定 D 链 6、7 位的鸟嘌呤，1109Ser 与 1107Lys 稳定 C 链的磷酸糖，这个蛋白质对 DNA 局部的稳定作用时 PAM 识别的关键。在特定的 DNA 序列处，氨基酸对 DNA 的稳定作用克服了 DNA 双链的稳定性并促进 DNA 双链解旋，使得 PAM 旁的 DNA 序列能够与 gRNA 的互补序列配对，形成 Cas9/DNA/RNA 复合体，从而实现对特定序列的切割（图 1.3）。由于氢键是氨基酸与核酸识别的主要分子间相互作用，并且依据理论模型 SpCas9 识别不同的 PAM 也是识别位点氨基酸与不同碱基间氢键能量不同，因此在设计突变体时、应着重考虑氢键对识别特异性的影响[10]。

随着计算机的发展，我们可以通过计算设计特异识别不同 PAM 的 Cas9 蛋白。计算设计可分为基于序列的计算设计和基于结构的计算设计。由于氨基酸与碱基间的识别取决于分子间氢键、静电作用及范德华相互作用，而以上作用力又取决于原子的空间位置，并且已知的晶体结构中存在很多非经典氨基酸-核酸识别，所以基于序列的计算设计不能够准确地设计出有识别特异性的突变体[11，12]。我们使用 Rosetta 作为模拟软件，基于结构设计特异识别 PAM 的 SpCas9 突变体。Rosetta 3.0 由 David Baker 实验室开发，作为比较成熟的蛋白质建模软件，可以用来设计蛋白质/DNA 相互识别，模拟突变体蛋白质/DNA 界面的结构[13]。该算法已被成功应用于改造 TALEN 蛋白对 DNA 切割的特异性。由于长的、极性氨基酸侧链的作用在模拟中较高的离散程度，使得在计算机模拟的过程中难以准确地发现蛋白质与 DNA 间良好相互作用的氨基酸构象。开发者试图通过在算法中提高突变体评估的准确性、调整能量函数、优化结构模拟的方法改善蛋白质与 DNA 相互作用的计算，但是仍不能准确地模拟氨基酸侧链的构象 [14，15, 16]。因此，在实际应用时仍需要根据具体情况调整计算的策略。

本文希望通过计算结构生物学的方法设计特异识别 PAM 的 SpCas9 突变体。首先，我们构建了的模拟与筛选 SpCas9 突变体的方法；然后，通过实验筛选的突变体验证了该方法的有效性；最后，根据建立的模拟方法设计识别新的 PAM 的 SpCas9 突变体，并发现了可能改变识别特异性的突变。

二、材料与方法

2.1 数据来源

2.1.1 SpCas9 结构数据

蛋白质结构和序列：SpCas9（PDB code：4un3），VQR 突变体（PDB code：5b2r），EQR 突变体（PDB code：5b2s），VRER 突变体（PDB code：5b2t）。以上数据均来自蛋白质数据库 www.rcsb.org。

2.1.2 氨基酸-核酸基序数据

氨基酸-核酸基序数据来自 AANT：amino acid-nucleotide interaction database。曾用网址 http://aant.icmb.utexas.edu。

2.2 模拟软件

2.2.1 计算结构生物学软件

蛋白质突变设计及分析软件：Rosetta 3.0 。（www.rosettacommons.org）

2.2.2 可视化和作图软件

Pymol （http://www.pymol.org）

2.2.3 脚本编写软件

Python 2.7

2.3 模拟与筛选方法

2.3.1 突变体序列的设计

蛋白质对 DNA 的识别可以分为氨基酸与核酸的直接识别和间接识别。直接识别指氨基酸直接与被识别核酸相互作用；间接识别指氨基酸在蛋白质/DNA 界面的、不直接作用于被识别的核酸，但通过影响识别相互识别的氨基酸与核酸的空间位置影响识别特异性[17]。氨基酸-核酸相互作用数据库（amino acid-nucleotide interaction database）中有从蛋白质结构数据库（www.rcsb.org）收集的氨基酸与核酸相互作用的数据（图 2.1），每一组氨基酸与其识别核酸的空间位置被称为一个基序[18]。基序数据可以作为直接识别的数据来设计蛋白质与 DNA 的直接识别突变。

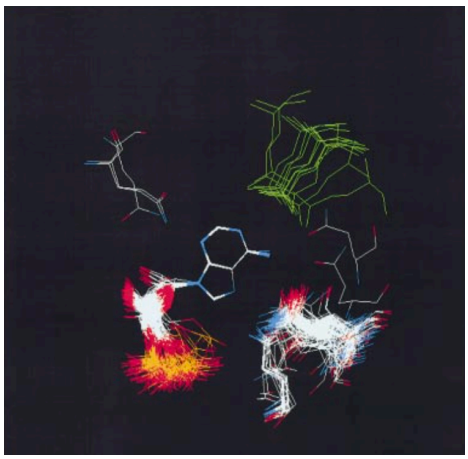

图 2.1. 基序示意图。在氨基酸核算数据库中，氨基酸对核酸的识别按照核酸种类整理分类。（Hoffman, M.M, 2004）

直接识别的突变在本课题中就是 SpCas9 蛋白 B 链 1333 位和 1335 位的氨基酸突变，所识别的 DNA 碱基分别是 D 链 6 位和 7 位核酸的碱基，也就是 PAM 的第二位和第三位核酸。我们以氨基酸-核酸相互作用数据库中氨基酸与碱基识别的基序数据为模版，使用 pymol 中的 align 功能，将氨基酸-核算基序的核酸与复合体 PAM 相应的核酸 align，计算基序中氨基酸的 N, C_α, and C 三个原子与 1333 位或 1335 位氨基酸的 N, C_α, and C 三个原子距离的 RMSD（root-mean-square deviation，均方根偏差）值。取 cut-off 值 3.0 Å，小于该值的氨基酸为候选直接识别氨基酸突变[19]。

间接识别的突变位点确定为距 DNA 中 D 链 6 位和 7 位核酸及 C 链配对的核酸 3.0 Å 范围内有交集的氨基酸，共 14 个，pdb 编号为 B 链的 1107、1109、1111、1135、1136、1139、1165、1215、1216、1218、1219、1221、1337 和 1339。同时再考虑直接识别位点突变氨基酸前后各一个的氨基酸。间接识别氨基酸的突变种类考虑全部 19 种非野生的突变，在具体计算时，第一轮只引入单位点突变，第二轮在第一轮的阳性基础上增加一个位点或将第一轮的阳性结果杂交。

2.3.2 突变体的构建

突变体结构的构建基于野生型 SpCas9 的晶体结构（4un3），通过 Rosetta 实现。核酸的突变通过 Rosetta 中 protein-DNA 设计的<DnaInterfacePacker>模块实现，氨基酸的突变通过<PackRotamersMovers>实现，突变氨基酸的旋转异构体选

择与野生型相近的构象。在突变体构建过程中不涉及突变碱基或氨基酸的结构优化。

2.3.3 突变体的结构模拟

突变体的结构通过 Rosetta 的 protein-DNA 设计中的<DesignProteinBackboneAroundDNA>模块实现，该模块通过 Monte Carlo 模拟优化蛋白质-DNA 界面氨基酸和核酸的构象，由于该模块没有识别构象是否稳定的算法，因此需要预设 Monte Carlo 模拟的步长。根据对步长与蛋白质-DNA 结合能量模拟结果的分析，10 步以后结合能就已接近平衡值（图 2.2），在模拟过程中我们设定步长统一为 20 步。

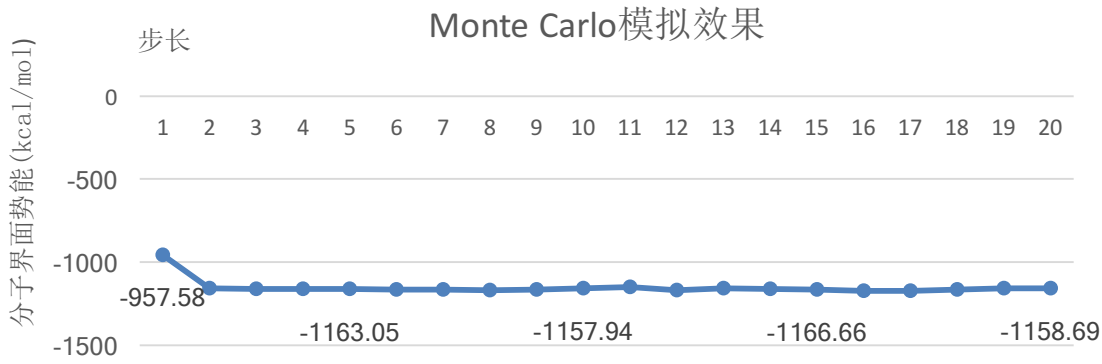

图 2.2. 构象优化过程中的步长-能量关系图。

进行完蛋白质 DNA 界面能量的整体优化后，我们再对识别位点 1333 和 1335 位的氨基酸进行能量最低的构象优化，利用 Rosetta 的<RotamerTrailsMinMover>模块实现。

2.3.4 评分函数的检验

由于 Rosetta 内置的评分函数未能筛选出实验已验证的突变体，根据理论模型，我们定义了氢键偏好的评分函数。评分函数计算识别位点氨基酸与 PAM 碱基间的氢键数目及能量。

氢键偏好的评分函数：PAM 碱基与识别氨基酸间（氢键数目，氢键能量）

以野生型 SpCas9 为初始模版，检验实验中发现识别 5’-NGA(G)-3’的突变体是否能在计算模拟被筛选出。突变位点限定在 1135、1335 和 1337 位，构建全部 8000 种突变体，进行结构优化后，利用氢键偏好的评分函数进行排序，对别模拟结果与实验结果。以类似方法验证识别 5’-NGCG-3’的突变体，但由于实验中有四个位点突变，为了降低计算量我们只考虑 1218、1335 和 1337 位极性或带电氨基酸的突变，1135 位置考虑突变为 Val 或 Asp。

2.3.5 突变体的设计计算流程

首先，根据基序设计的结果确定直接识别位点的氨基酸，再引入一个间接识别突变；然后，构建突变体，并优化突变体结构；最后，根据氢键偏好的评分函数筛选突变体。对其中的阳性模拟结果进行进一步的突变设计。

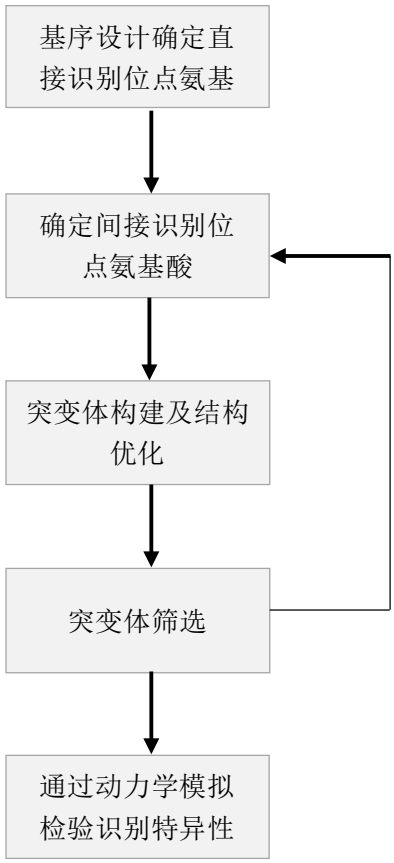

图 2.3. 特意识别 PAM 的 SpCas9 突变体的计算设计流程图。

三、研究结果

3.1 模拟与筛选方法的验证

为了通过已知的实验结果验证模拟的可靠性，我们构建了识别 5’-NGAG-3’的 8000 个突变体和识别 5’-NGCG-3’的 3456 个突变体。由于实验结果中识别 5’-NGAG-3’的突变体共有 3 个突变氨基酸（B 链的 1135、1335 和 1337），所以将突变位点限定在实验中的 3 位点，共有 8000 种突变体。我们模拟了这 8000 种突变体结构并通过 Rosetta 内置的评分函数和自定义的氢键偏好的评分函数筛选，发现实验筛选的突变体在氢键偏好的评分函数中更为突出。我们也利用了识别 5’-NGCG-3’的突变体验证模拟的可靠性，由于 4 个位点理论上共有 160000 种突变体，计算量过于庞大，因此我们只考虑这 1218、1335 和 1337 三个位点的带电或亲水氨基酸突变和 1135V 或 1135D 突变，共 2*12*12*12=3456 种突变体。我们模拟了这 3456 种突变体的结构，同样发现实验验证的突变体在氢键偏好的评分函数中能够突出。

3.1.1 Rosetta 内置评分函数筛选结果

Rosetta 内置的评分函数对每个突变体计算了一个 fitness 值，值越小代表识别特异性越强。在 Rosetta 的评分结果中，我们发现不同突变体间的区分度不够明显（图 3.1），并且实验中已经验证活性的突变体评分普遍较高，都在 8000 个突变体的后 50%（表 3.1）。

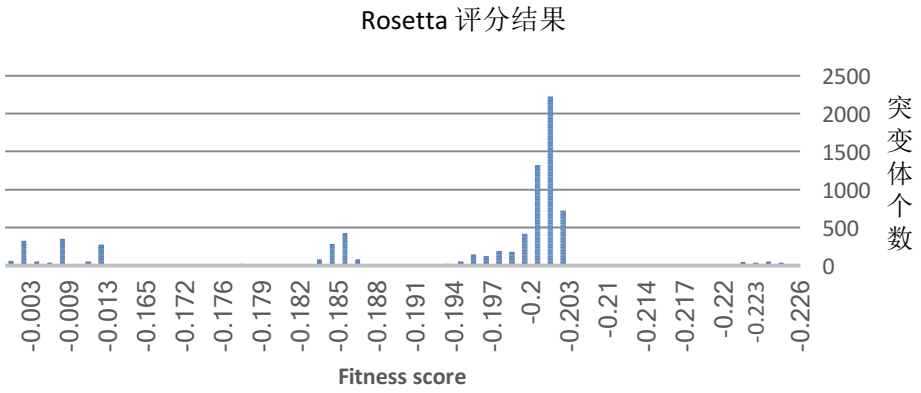

图 3.1. Rosetta 内置评分函数对识别 5’-NGAG-3’的 8000 个突变体评分结果。

表 3.1. 实验验证的突变体 fitness 值及排名

突变体	FITNESS	RANK
DRR	-0.200	54%
VRR	-0.199	58%
YRR	-0.183	80%
NRR	-0.200	54%
ERR	-0.200	54%
DQR	-0.200	54%
VQR	-0.200	54%
YQR	-0.184	79%
NQR	-0.200	54%
EQR	-0.200	54%

3.1.2 氢键偏好的评分函数

我们将识别 5’-NGAG-3’的 8000 个突变体模拟结果使用氢键偏好的评分函数排序（图 3.2）。结果显示实验中验证的突变体评分明显优于 Rosetta 内置的评分函数，实验中验证的识别特异性较强并已解析晶体结构的 VQR、EQR 突变体在氢键偏好的评分函数下能够明显地被筛选出来（表 3.2）。

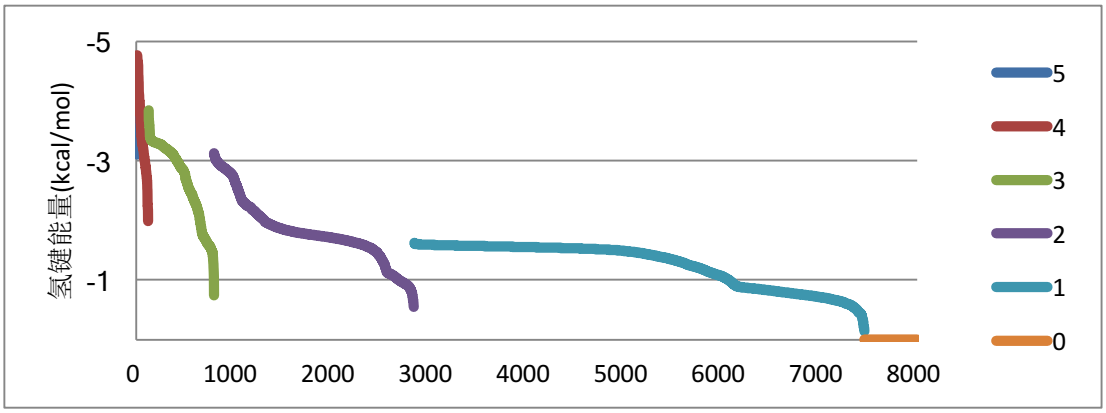

图 3.2. 氢键偏好评分函数对识别 5’-NGAG-3’的 8000 个突变体评分结果，不同颜色代表不同氢键数目。

表 3.2. 实验验证的突变体氢键偏好评分及排名

突变体	SCORE	RANK
DRR	4, -2.43	1.5%
VRR	3, -2.67	6.4%
YRR	3, -3.30	2.3%
NRR	4, -3.29	0.8%
ERR	4, -4.47	0.2%
DQR	2, -3.00	10.4%

VQR	4, -3.89	0.5%
YQR	3, -3.66	1.6%
NQR	2, -2.87	11.6%
EQR	4, -3.98	0.5%

我们也将识别 5’-NGCG-3’的 3456 个突变通过氢键偏好的评分函数排名，发现实验筛选的突变体依旧突出（表 3.3）。

表 3.3. 识别 5’-NGCG-3’的突变体氢键偏好评分及排名

	VGER	DRER	VRER
SCORE	1，-0.55	3，-2.02	4，-3.64
RANK	76.4%	1.1%	0.1%

根据对已知突变体的计算模拟，可以证明我们的模拟方法和氢键偏好的评分函数的有效性。

3.2 直接识别设计结果

直接识别设计分为 PAM 单位点核酸替换的直接识别设计和 PAM 双位点核酸替换的直接识别设计。两种设计都根据基序模拟的结果确定，如果基序数据中某种已知的氨基酸-核酸识别符合 SpCas9 对 PAM 识别时氨基酸-核酸的空间位置，则尝试在后续设计中模拟这种直接识别的突变体。

3.2.1 基序模拟结果

根据氨基酸-核酸相互作用数据库的基序数据，我们计算了 PAM 序列第二位或第三位碱基为非野生型（5’-NGG-3’）时，所有基序数据中的氨基酸与 SpCas9 识别位点氨基酸（1333 或 1335）N, C_α, and C 三个原子的 rmsd 值，并取最小值（表 3.4、3.5）。最小值越小说明已知的氨基酸-核算识别方式越符合 SpCas9 对 PAM 的识别，则对这种识别方式进行后续模拟成功的可能性越大。

表 3.4. PAM 第二位核酸的基序模拟结果

A		T		C	
氨基酸	最小 rmsd/ Å	氨基酸	最小 rmsd/ Å	氨基酸	最小 rmsd/ Å
Arg	1.49	Arg	1.01	Arg	10.43
Asn	3.41	Asn	2.36	Asn	3.03
Asp	3.59	Asp	9.97	Asp	2.46
Cys	5.49	Cys	4.15	Cys	3.50
Gln	2.01	Gln	2.16	Gln	3.11
Glu	5.10	Glu	13.58	Glu	2.11
His	6.07	His	2.46	His	15.42
Lys	2.56	Lys	2.29	Lys	13.85
Met	6.02	Met	\	Met	5.91
Ser	4.59	Ser	4.79	Ser	3.89
Thr	4.06	Thr	7.71	Thr	3.81
Trp	19.50	Trp	\	Trp	20.18
Tyr	2.18	Tyr	2.76	Tyr	2.25

表 3.5. PAM 第三位核酸的基序模拟结果

A		T		C	
氨基酸	最小 rmsd	氨基酸	最小 rmsd	氨基酸	最小 rmsd
Arg	4.13	Arg	2.47	Arg	11.05
Asn	4.48	Asn	2.60	Asn	3.32
Asp	10.16	Asp	10.72	Asp	2.93
Cys	11.24	Cys	4.38	Cys	3.86
Gln	9.16	Gln	2.26	Gln	2.38
Glu	9.56	Glu	13.72	Glu	1.94
His	11.02	His	2.91	His	16.59
Lys	4.13	Lys	1.09	Lys	14.32
Met	10.97	Met	\	Met	3.88
Ser	5.91	Ser	5.29	Ser	4.06
Thr	8.48	Thr	7.92	Thr	3.98
Trp	15.66	Trp	\	Trp	20.86
Tyr	9.26	Tyr	2.14	Tyr	1.92

我们取 3.0Å 的临界值，确定识别 PAM 第 2、3 位不同核酸时的氨基酸（表 3.6）。

表 3.6. PAM 核酸的基序设计结果

碱基位置	目标核酸	识别氨基酸
PAM 第二位核酸	A	Gln，Lys，Tyr
（D 链 6 位核酸）	C	Asp，Glu，Tyr
	T	Gln，Asn，His，Lys，Tyr
PAM 第三位核酸	C	Asp，Gln，Glu，Tyr
（D 链 7 位核酸）	T	Asn，Gln，His，Lys，Tyr

3.2.2 直接识别设计结果

PAM 单核酸替换的突变体直接识别设计即基序计算结果（表 3.7）。除去单核酸替换的 PAM 设计，我们也尝试了设计双位点的核酸替换，在设计双位点核酸替换时，由于 1333 位氨基酸与 1335 位氨基酸处于识别 PMA 的构象时，两个 β 碳原子的距离仅有 3Å，因此为了使氨基酸与碱基间的识别稳定，识别位点的两个氨基酸最好避免相互之间形成氢键，所以在设计时我们只考虑在识别位点有相同电负性侧链的氨基酸。符合侧链具有相同电负性，并且侧链长度适合的氨基酸有 Lys、Tyr、Glu。当 1333 与 1335 位氨基酸种类确定时，我们根据之前基序计算的结果反向确定识别位点的氨基酸（表 3.8）。

表 3.7. PAM 单核酸替换的 SpCas9 突变体

	目标 PAM	1333 位氨基酸	1335 位氨基酸
单位点碱基替换	5’-NAG-3’	Gln、Lys 、Tyr	\
	5’-NCG-3’	Asp、Glu、Tyr	\
	5’-NTG-3’	Gln、Asn、His、Lys、Tyr	\
	5’-NGC-3’	\	Asp、Gln、Glu、Tyr
	5’-NGT-3’	\	Asn、Gln、His、Lys、Tyr

表 3.8. PAM 双核酸替换的 SpCas9 突变体

	识别位点氨基酸	识别 PAM 序列
双位点碱基替换	Lys-Lys	5’-NTT-3’，5’-NAT-3’
	Glu-Glu	5’-NCC-3’
	Tyr-Tyr	5’-NAC-3’，5’-NAT-3’，5’-NCC-3’，5’-NCT-3’，5’-NTC-3’，5’-NTT-3’

3.3 单位点碱基替换的突变体设计

3.3.1 突变体初步筛选结果

根据基序设计的结果，我们确定了直接识别的设计，即识别特定 PAM 时识别位点氨基酸的种类。对于间接识别，我们选择突变位点为距 PAM 中 D 链 6 位和 7 位核酸及 C 链配对的氨基酸 3.0 Å 的氨基酸，共 14 个，pdb 编号为 1107、1109、1111、1135、1136、1139、1165、1215、1216、1218、1219、1221、1337 和 1339。对于 PAM 第一个碱基替换的突变体（D 链第 6 位核酸），我们还考虑 1333 位氨基酸前后的两个氨基酸，1332 和 1334；对于 PAM 第二个碱基替换的突变体（D 链第 7 位核酸），我们还考虑 1335 位氨基酸前后的两个氨基酸，1334 和 1336。

我们对每种直接识别设计的结果引入一个间接识别突变，再进行结构模拟。模拟结果中若突变的的直接识别氨基酸与目标碱基间有氢键形成则为阳性结果，若没有氢键或未突变的直接识别氨基酸与碱基间氢键被破坏则为阴性结果。

我们发现 PAM 第二位碱基替换的突变体几乎没有阳性模拟结果（表 3.9），识别 5’-NCG-3’的 1334Trp 突变体结果为阳性，但由于氢键太弱，故放弃这条路线的突变体设计，不再尝试仅替换 PAM 第二位的碱基。PAM 第三位碱基替换的突变体设计结果明显好于第二位碱基替换的突变体设计（表 3.10）。比较好的结果集中于 1335 位突变为 Glu 时识别 5’-NGC-3’ PAM 和 1335 位突变为 Lys 时识别 5’-NGT-3’ PAM。我们对识别这两种 PAM 的突变体进行了进一步设计。

表 3.9 PAM 第二位碱基替换的突变体设计结果

目标 PAM	1333 位氨基酸突变	阳性突变
NAG	Gln	NONE
	Lys	NONE
	Tyr	NONE
NCG	Glu	1334Trp
	Tyr	NONE
NTG	Gln	NONE
	Asn	NONE
	His	NONE
	Lys	NONE
	Tyr	NONE

表 3.10. PAM 第三位碱基替换的突变体设计：

目标 PAM	1335 位氨基酸突变	阳性突变
NGC	Asp	NONE
	Gln	NONE
	Glu	1107-P，1111-T，1218-W，1334-I/V，1336-F，1337-F
	Tyr	NONE
NGT	Asn	NONE
	Gln	1107-S，1221-F
	His	NONE
	Lys	1107-I，1109-C/P，1111-G/M，1135-S/P/F/R，1136-I/M/F/W，1139-C/F/W 1165-E/N/L/W/V，1216-M/F/V，1218-H/P/L，1219-E/G/P/M/W，1336-R，1337-E/P
	Tyr	NONE

3.3.2 识别 5’-NGC-3’的突变体进一步设计结果（Glu）

在间接识别突变的结果中，当目标序列为 5’-NGC-3’，直接识别突变为 Glu 时，有 7 个阳性的突变结果。我们分别以这七个突变体为初始结构，继续在间接识别位点引入单碱基突变。在阳性结果中，我们将提升识别特异性的突变杂交，发现 K1107P、D1135V、Q1221H、K1334L 几个突变对识别特异性非常重要。

我们尝试了对 PVHLE（K1107P、D1135V、Q1221H、K1334L、R1335E）突变体（图 3.3）进行分子动力学模拟，检验识别位点氨基酸与碱基间氢键的稳定性。结果表明 1335 位 Glu 对目标 7C 的识别非常稳定，有较强、较稳定的氢键，

但 1333 位 Arg 对 6G 的识别不稳定。模拟开始时 Arg 与 6G 间还保持有稳定的氢键，2.5ns 后氢键便被破坏（图 3.4）。根据对动力学模拟结果的仔细分析，我们发现 Arg 与 6G 间氢键的破坏与 1335 位 Glu 有直接关系。Glu 的羰基集团带有负电而 Arg 的氨基带有正点，Glu 与 Arg 在平衡位置附近摆动时，相互之间极易形成氢键，使 Arg 的侧链旋转，偏离与 6G 识别的位置，并在后续的模拟过程中被周围的电负性为负的基团吸引，不再与 6G 识别。

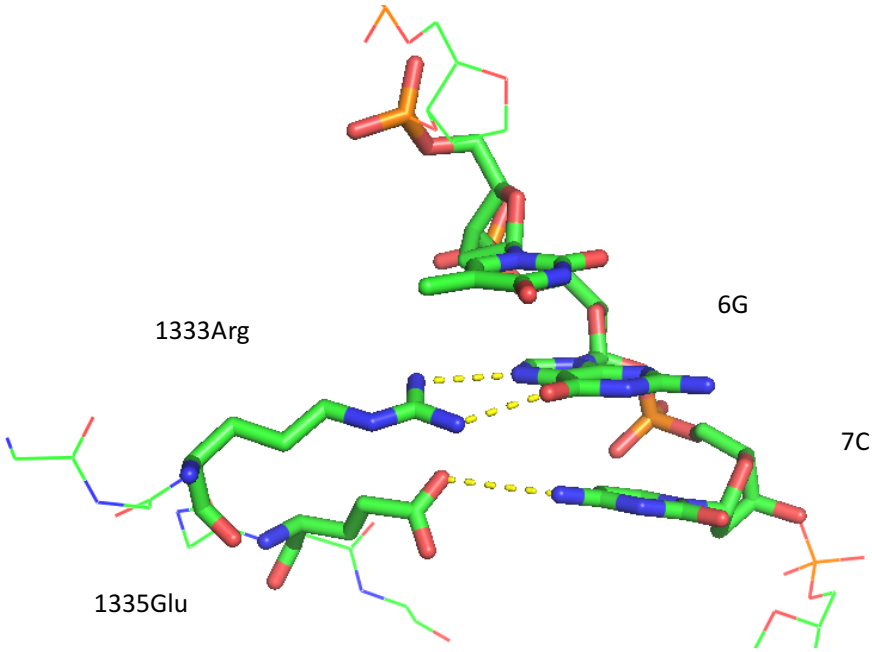

图 3.3. PVHLE 突变体识别 5’-NGC-3’。1333Arg 与 6G 形成 2 个氢键，1335Glu 与 7C 形成 1 个氢键，K1107P、D1135V、Q1221H、K1334L 未在图中标出。

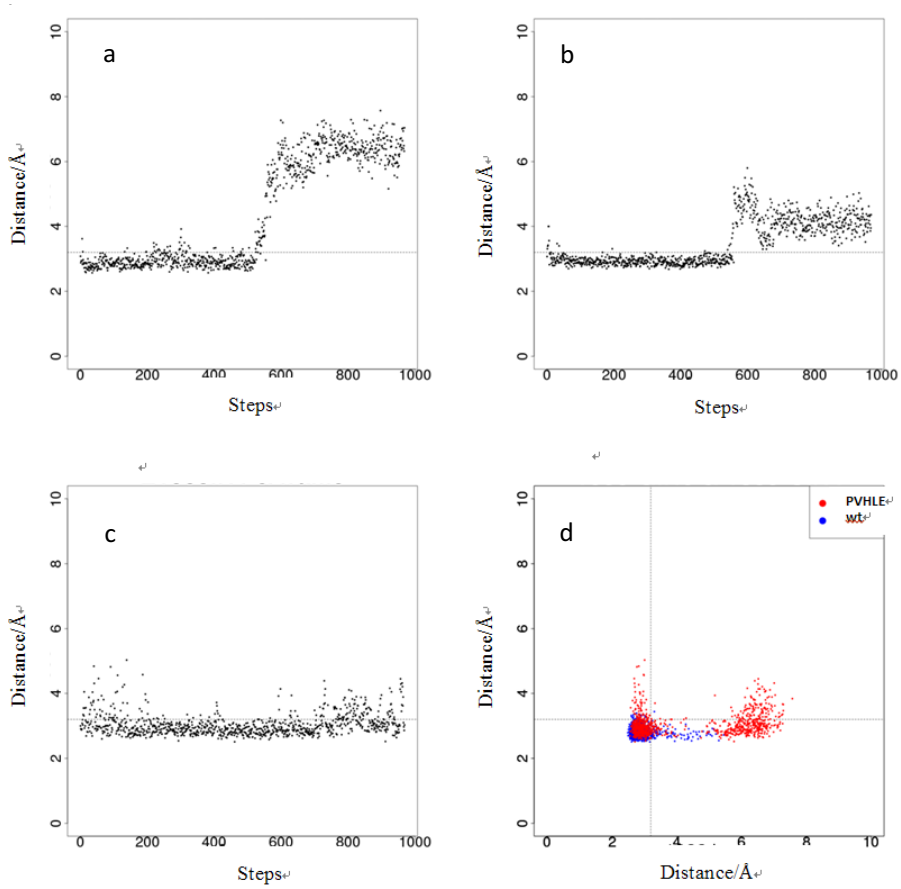

图 3.4. PVHLE 突变体动力学模拟分析结果。1333 位 Arg 的氨基与碱基 N6 的原子间距离（a）；1333 位 Arg 的氨基与碱基 N7 的原子间距离（b）；1335 位 Glu 的羰基与碱基 N4 的原子间距离（c）；野生型 SpCas9 与 PVHLE 突变体氢键稳定性的对比（d）。

为了使 1335 位 Glu 在与 7C 稳定识别的同时，不会破坏 1333 位 Arg 对 6G 的识别，我们希望设计的突变体 1335 位 Glu 的羰基和羧基一个与 7D 形成氢键，另一个基团通过氢键被稳定。根据晶体结构，在 1337 为引入突变有较大的可能实现设计目的，经过模拟，我们发现 PVHEH（K1107P、D1135V、Q1221H、R1335E、T1337H）突变体可能比较符合目标，在 1337 位引入一个 His 突变，通过 1337His 的咪唑与 1335Glu 的相互作用防止 1335Glu 与 1333Arg 形成氢键。但动力学模拟的结果 1333Arg 还是偏离了与 6G 的识别位置。

3.3.3 识别 5’-NGC-3’的突变体进一步设计结果（Lys）

第一轮模拟的突变体中，1335 位为 Lys 且识别 5’-NGC-3’的突变体较多，有 1107I、1109C/P、1111G/M、1135S/P/F/R、1136I/M/F/W、1139C/F/W、1165E/N/L/W/V、

1216M/F/V、1218H/P/L、1219E/G/P/M/W、1336R、1337E/P。我们直接将第一轮的阳性突变杂交，获取双位点间接识别突变的突变体。结果中，630 个突变体有 168 个阳性结果，30 种单位点突变对识别特异性的提升没有明显的差异。这个方向的设计需进一步研究后确定。

3.4 双位点碱基替换的突变体设计

双位点碱基替换的突变体设计中（表 3.11），当识别位点氨基酸为 1333Glu 和 1335Glu 时，K1334W 突变能够识别 5’-NCC-3’。除了 1334W 突变，在其它 16 个间接识别位点再引入单突变的 304 个突变体有 244 个有对 5’-NCC-3’的识别特异性。我们初步判断 EWE（1333E、1334W、1335W）突变体有 5’-NCC-3’的识别特异性。可以通过动力学模拟或实验进行进一步验证.

表 3.11. 双位点碱基替换的突变体设计结果

直接识别氨基酸	目标 PAM 序列	阳性结果
LYS-LYS	5’-NTT-3’、5’-NAA-3’	NONE
GLU-GLU	5’-NCC-3’	1334-W
TYR-TYR	5’-NAC-3’，5’-NAT-3’，5’-NCC-3’，5’-NCT-3’，5’-NTC-3’，5’-NTT-3’	NONE

3.5 结论

根据 PAM 序列对于 SpCas9 特异性靶向、切割 DNA 的理论模型，我们确定采用自定义的、氢键偏好的评分函数来筛选突变体，并根据实验获得的识别 5’-NGAG-3’、5’-NGCG-3’PAM 的突变体验证了该评分函数的有效性。在设计特异识别 PAM 的 SpCas9 突变体时，我们根据氨基酸-核酸相互作用库的基序数据设计了直接识别位点的氨基酸，在直接识别确定的基础上，我们模拟了 PAM 周围间接识别氨基酸的突变，并发现了一系列能够提升识别 5’-NGC-3’、5’-NGT-3’和 5’-NCC-3’特异性的突变位点。对识别 5’-NGC-3’的 PVHLE 突变体，我们还通过动力学模拟进一步分析了其识别特异性，但发现识别并不稳定。本研究的结果可以为进一步设计特异识别 PAM 的 Cas9 突变体提供参考。

四、讨 论

4.1 蛋白质/DNA 识别的设计

蛋白质/DNA 相互识别的设计可分为基于序列的设计或基于结构的设计，本质上是基于已知识别的设计和基于分子间势能的设计。

基于分子间势能的计算设计优点在于物理原理上的正确性。由于原子的空间或静电冲突，蛋白质和 DNA 间的势能对氨基酸的序列和原子的空间位置非常敏感。准确的模拟由于序列的改变而导致的分子间势能的改变将决定设计的成功与否，但目前还没有非常有效的模拟方法[20]。基于已知识别的计算设计优点在于操作上的可行性。但由于已知的识别都来自于实验条件获得的晶体结构信息，我们不能模拟试验中未得到的识别信息。

本课题结合了两种方法来设计特异识别 PAM 的 SpCas9 突变体， 基于已知识别的直接识别设计和基于分子间势能的间接识别设计。氨基酸-核算相互识别数据库的基序数据属于已知的氨基酸-DNA 相互识别，我们通过该数据确定了直接识别位点的氨基酸种类。但由于数据来源的基序周围原子环境不同，我们考虑了在 PAM 周围的氨基酸引入突变，并通过 Rosetta 的优化突变体 PAM 周围蛋白质/DNA 界面的结构，即降低分子界面的势能。这样，我们实现了降低计算量的情况下，保证了模拟的准确性。

4.2 本课题计算的理论模型及评分函数

本课题的理论模型可以概括为，SpCas9/RNA 复合体识别 PAM，促进了 DNA 的解旋及 SpCas9/RNA/DNA 复合体的形成，SpCas9 对 PAM 识别的特异性通过 1333 位和 1335 位氨基酸与 DNA D 链 6、7 位碱基间的氢键体现。但这些假设都是基于晶体结构的分析得出的。在 PAM 识别过程中复合体发生了构象变化，我们假设氨基酸与特定碱基间的氢键是降低构象变化的活化能的重要因素。我们计算时使用的 SpCas9/RNA/DNA 复合体的构象不是复合体构象变化瞬间的构象，因此不能准确地从分子间势能的角度体现识别特异性，这也是本课题的主要缺陷所在。但由于目前还没有模拟过渡态构象的技术，计算设计只能用晶体构象近似处理。

氨基酸与碱基分子间的作用除了氢键还包括静电相互作用和范德华相互作用，我们的评分函数只考虑氢键也是一种近似处理。特定的氨基酸序列与不同核酸序列间的势能区别很大程度上取决于氢键。通过对实验已经验证突变体的模拟，我们发现氢键是最好的筛选方法。

在对模拟结构的分析中，我们发现有很多结果中氨基酸与碱基间的氢键不是1333 位氨基酸与 D 链 6 位核酸、1335 位氨基酸和 D 链 7 位核酸的对应关系，1333 位氨基酸可以与 C 链 6 位或者 7 位的核酸碱基形成氢键，1335 位氨基酸可以与 C 链 5 位或者 6 位的氨基酸形成氢键。这些结果的数量与野生型识别方式的数量相近。由于无法验证这些氢键分布是否也与识别特异性有相关性，我们在本次模拟中将这些结果定义为阴性模拟结果。但考虑到 SpCas9 对 PAM 的识别一直存在一定的容错率，说明由 PAM 引起的构象变化在能量有一定的诱发范围，非典型氨基酸-碱基识别也许也可以满足能量的需求[21]。因此，对于模拟结果中非典型的氨基酸-碱基识别需要试验验证，若能获得与模拟结果相同的晶体结构，不仅能够促进人们对 SpCas9 催化机理的认识，还可以极大增加计算设计的筛选范围，提高计算设计成功的可能[22]。

参考文献

1. Garneau, J. E. et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature 468, 67–71 (2010).
2. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337, 816–821 (2012).
3. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature 507, 62–67 (2014).
4. Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology. Nature Methods 10, 957–963 (2013).
5. Bikard, D. et al. Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. Nucleic Acids Res. 41, 7429–7437 (2013).
6. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nature Biotech nol. 31, 233–239 (2013).
7. Gilbert,L.A.et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell 154, 442–451 (2013).
8. Kleinstiver, B.P., et al., Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature, 2015. 523(7561): p. 481-5.
9. Hirano S, Nishimasu H, Ishitani R, Nureki O. Structural basis for the altered PAM specificities of engineered CRISPR-Cas9. Mol Cell. 2016;61:886–94.
10. Anders, C., et al., Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature, 2014. 513(7519): p. 569-73.
11. Ashworth J., et al., Computational reprogramming of homing endonuclease specificity at multiple adjunct base pair. Nucleic Acid Res, (2010), No. 16; 5601-5608.
12. Luscombe NM, Laskowski RA, Thornton JM. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res. 2001;29:2860–2874.
13. Pabo CO, Nekludova L. Geometric analysis and comparison of protein–DNA interfaces: why is there no simple code for recognition? J Mol Biol 2000;301:597–624.
14. Fleishman SJ, Leaver-Fay A, Corn JE, Strauch E-M, Khare SD, et al. (2011) RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. PLoS ONE 6(6): e20161.
15. Havranek JJ, Duarte CM, Baker D. A simple physical model for the prediction and design of protein–DNA interactions. J Mol Biol 2004;344:59–70.
16. Ashworth, J., et al., Computational redesign of endonuclease DNA binding cleavage specificity. Nature, 2006. 441(7093): 656-659.

17. Joyce AP, Zhang C, Bradley P, Havranek JJ. Structure-based modeling of protein: DNA specificity. Briefings in Functional Genomics. 2015;14(1):39-49. doi:10.1093/bfgp/elu044.
18. Hoffman MM, Khrapov MA, Cox JC, Yao J, Tong L, Ellington AD (2004) AANT: the amino acid-nucleotide interaction database. Nucleic Acids Res 32:D174–D181.
19. Havranek JJ, Baker D. Motif-directed flexible backbone design of functional interactions. Protein Sci 2009;18: 1293–305
20. Dey S, Pal A, Guharoy M, Sonavane S, Chakrabarti P. Characterization and prediction of the binding site in DNA-binding proteins: improvement of accuracy by combining residue composition, evolutionary conservation and structural parameters. Nucleic Acids Res. 2012;40:7150–7161.
21. Anders, C., Bargsten. K., Jinek. M., Structural Plasticity of PAM Recognition by Engineered Variants of the RNA-Guided Endonuclease Cas9. Molecular Cell 61, 895-902 (2016).
22. O'Geen, H., Yu, A.S., Segal, D.J. How specific is CRISPR/Cas9 really? *Current Opinion in Chemical Biology* (2015), 29, pp. 72-78.

致谢

首先要感谢黄老师两年来对我的课题研究指导。其次，感谢王国栋师兄和姚睿捷对我计算机方面的帮助，感谢袁慧师兄和汤洪海师兄在实验技术方面对我的指导，同时也感谢实验室其他师兄师姐对我的指导和帮助。最后感谢父母对我从事生物学方向研究的支持。