通过蛋白设计增加 SpCas9 的 PAM 序列兼容性

生命科学学院　王亦帆
指导教师　黄强

摘要：CRISPR-Cas9 是一种 RNA 引导的核酸内切酶。由于它具有高度可编程性，已经成为基因编辑领域的重要工具酶。除了引导 RNA 外，Cas9 必须识别一段特殊的序列，称作 protospacer adjacent motif （PAM）。由于 PAM 的存在，Cas9 无法识别任意的 DNA 序列，因此其靶向空间受限。对于最常用的 *Streptococcus pyogenes* Cas9（SpCas9），通过实验手段，已经开发出了多种突变体，可以识别不同的 PAM 序列。本文试图引入计算机辅助设计方法来改造 SpCas9 的 PAM 序列兼容性。通过基于结构的自由能计算，本文开发了一种预测 SpCas9 及其突变体 PAM 兼容性的物理模型，可以由突变体的氨基酸序列预测其 PAM 兼容性。基于此模型，我们试图构造一种新型的 SpCas9 突变体，以获得更强的 PAM 兼容性。虽然该突变体的活性检测以失败告终，但是它启发我们进一步改进 PAM 兼容性预测模型。经过改良的 PAM 兼容性预测模型采用了浅层神经网络，引入更多参数，模型的准确度有了很大的提升。我们相信计算机辅助蛋白设计将成为 Cas9 突变体筛选的重要手段。
关键词：CRISPR-Cas9　PAM　计算机辅助蛋白设计

Abstract: CRISPR-Cas9 is an RNA-guided DNA endonuclease. Since it is highly programmable, it has been widely applied as a tool enzyme in gene editing. Cas9 also recognize a short DNA sequence called protospacer adjacent motif (PAM) other than guide RNA, and thus its target space is limited. For the most widely used Streptococcus pyogenes Cas9 (SpCas9), some variants have been developed, which target different PAM sequence, by experimental methods. This article tried to introduce computational approach to redesign the PAM compatibility of SpCas9. Through structure based free energy calculation, we developed a physical model to predict PAM specificity of SpCas9 and its variants, with amino acid sequences. Based on this model, we tried to develop a new SpCas9 variant. Although this new variant was proofed to be a failure, it inspired us to improve the PAM specificity prediction model. The new model used one-hidden-layer neural network and adopted more parameters, and its accuracy was greatly improved. We believe this computational protein designing approach will become an important method in development of Cas9 variant.
Keywords: CRISPR-Cas9, PAM, computational protein design

引言

CRISPR-Cas（clustered regularly interspaced short palindromic repeats / CRISPR-associated proteins）系统是细菌和古生菌中发现的一种针对外源核酸的获得性免疫系统。该系统可以通过 CRISPR 适应、CRISPR-RNA（crRNA）表达和 crRNA 干扰的过程来实现获得性免疫功能。具体来说，细菌或古生菌可以整合外源 DNA 并以此为模板生产 crRNA，最终通过 crRNA 来靶向降解外源 DNA。[1]其中，CRISPR-Cas9 系统是一种 II 型 CRISPR-Cas 核酸酶，负责 crRNA 干扰过程。它由 Cas9 蛋白、crRNA 和 tracrRNA（trans-activating crRNA）构成，通过 crRNA 和靶 DNA 的互补配对，来靶向切割目的

[1] Sorek et al. CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea. Annual Review of Biochemistry. 2013.

DNA。[2]由于它具有高度的可编程性，通过设计 crRNA 即可靶向目的 DNA 序列，因此在基因工程和基因编辑领域具有巨大的应用价值。目前应用最广泛的是化脓性链球菌（Streptococcus pyogenes）的 Cas9 系统（SpCas9）。工程化的 SpCas9 使用 single guide RNA（sgRNA）取代原本的 crRNA 和 tracrRNA[3]，并已证实可以在多种生物细胞的基因组中进行编辑操作。除了 Cas9 介导的 DNA 双链断裂（DSB）外，经过改造的 Cas9 还可以进行碱基编辑[4]和 prime editing[5]，从而做到碱基的转换、颠换，以及插入或缺失突变。因此，Cas9 介导的基因编辑技术有望成为各种人类遗传疾病的终极解决方案，并且在基因工程的各个领域具有广阔的应用前景。

CRISPR-Cas9 系统主要存在三种缺陷。其一是 Cas9 的小型化问题。SpCas9 蛋白全长超过 1300aa，将 SpCas9 与其 sgRNA 基因包装进入病毒载体十分困难，因此有必要设计更为小型化的突变体，或者寻找更小的 Cas9 同源物。其二是脱靶问题。Cas9 有概率靶向与 sgRNA 不完全匹配的 DNA，因而会造成一些未知的突变，这些脱靶造成的突变在实际应用过程中会带来潜在的风险。其三是 PAM 序列兼容性的限制。除了 sgRNA 之外，由于 Cas9 蛋白和靶 DNA 的特异性相互作用，Cas9 还必须识别一段 PAM（protospacer adjacent motif）序列。[6]相对于 sgRNA 和靶 DNA 的碱基互补，Cas9 蛋白和靶 DNA 的相互作用更为复杂，针对 PAM 序列的特异性重新设计也更加困难。由于 PAM 的限制，Cas9 无法靶向任意 DNA 序列，或者说其靶向空间受到限制。对于 DNA 的精确操作，例如碱基编辑，PAM 序列的限制就更加明显。因此，有必要开发一种 PAM 序列兼容性更好的 Cas9 突变体。

通过 X 射线晶体学方法，SpCas9-sgRNA-DNA 的复合物结构已经得到解析。[7] [8]整个 SpCas9 蛋白可以分为 REC 叶（recognition lobe）和 NUC 叶（nuclease lobe）两部分。在 PAM 序列上游处，双链 DNA 解旋，其中靶链（target strand，TS）与 sgRNA 杂交，结合在 REC 叶处，PAM 序列所在的非靶链（non-target strand，NTS）则结合在 NUC 叶处。HNH 结构域和 RuvC 结构域分别催化 TS 和 NTS 的断裂。在 PAM 序列处，PI 结构域则与 PAM 序列处双螺旋 DNA 直接接触，其中 1333 和 1335 位精氨酸残基伸入 DNA 大沟中，与 NTS 中 2 位和 3 位鸟嘌呤形成特异性氢键，因此 SpCas9 识别 NGG PAM。如果将 1333 和 1335 位精氨酸残基突变为丙氨酸，也会降低 SpCas9 的活性。[7]总之，PI 结构域的改造将是改变 PAM 特异性的关键。

在 SpCas9 的各个同源物中， PAM 序列并不保守，例如在 Streptococcus canis Cas9 (ScCas9)中，PAM 序列为 NNG[9]，而在 Staphylococcus aureus Cas9（SaCas9）中，PAM 序列为 NNGRRT（R=A 或 G）[10]。由此可见，PAM 序列的特异性具有改造的空间。通过实验手段，已经获得了多种 SpCas9 的突变体，识别不同的 PAM。例如，xCas9 可以识别 NGN、

---

2 Makarova et al. Evolution and classification of the CRISPR–Cas systems. Nature Reviews Microbiology. 2011.
3 Mali et al. RNA-Guided Human Genome Engineering via Cas9. Science. 2013.
4 Gaudelli et al. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature. 2017.
5 Anzalone et al. Search-and-replace genome editing without double-strand breaks or donor DNA. Nature. 2019.
6 Sternberg et al. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature. 2014.
7 Anders et al. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature. 2014.
8 Jiang et al. Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. Science. 2016.
9 Chatterjee et al. Minimal PAM specificity of a highly similar SpCas9 ortholog. Science Advances. 2018.
10 Nishimasu et al. Crystal structure of Staphylococcus aureus Cas9. Cell. 2015.

GAA、TGT 或 GAT PAM[11]，而 SpCas9-NG 可以识别 NG PAM[12]。

此外，通过计算结构生物学和实验相结合的方法，也可以进行 PAM 特异性的重新设计。通过基于结构的分子动力学模拟（molecular dynamics simulation）和自由能微扰（free-energy perturbation），得到了 PAM 兼容性更强的 SaCas9 突变体，SaCas9-NR 和 SaCas9-RL，它们可以识别 NNGRRN PAM。[13]该方法假设 SaCas9 与 PAM DNA 的结合自由能（ΔG）可以作为 PAM 兼容性的判据，因为由结合自由能可以预测 Cas9 与 DNA 的亲和力，结合自由能越低，则亲和力越强。通过自由能微扰，可以计算突变前后 SaCas9 与 DNA 的结合自由能改变（ΔΔG），然后筛选出ΔΔG 最小的突变体。ΔΔG 越小，则表明突变体可能与 DNA 有更强的亲和力。

然而，该方法也有一些缺陷。首先，该方法仅计算了固定 PAM 序列时的结合自由能改变。实际上，PAM 序列兼容性应当取决于各种不同 PAM 序列与 Cas9 的结合能的分布情况：结合能更低的 PAM 序列拥有更强亲和力，更有可能被兼容，反之，结合能高的 PAM 序列则可能无法被兼容。其次，基于分子动力学模拟的自由能微扰法需要消耗大量的计算资源，很难进行大批量计算。受此启发，我们构建了一种计算机辅助手段，用于扩展 SpCas9 的 PAM 序列兼容性。

1　Cas9 识别 PAM 的化学平衡模型

首先需要构建由氨基酸序列预测 SpCas9 及其突变体 PAM 兼容性的模型，随后即可以此模型为基础进行突变设计并筛选可能的突变体。我们通过同源建模、结合能计算和 PAM 识别过程模型，拟合现有实验数据，构建了 Cas9 识别 PAM 的化学平衡模型。

1．1 PAM 识别过程模型

已有实验表明，Cas9 靶向并切割 DNA 的过程分为如下几步：首先，游离的 Cas9-sgRNA 复合物移动到 DNA 附近，并且尝试与 DNA 结合。当 Cas9 寻找到合适的 PAM 序列时，则会结合到 DNA 上，并尝试使 DNA 解旋，其中 TS 与 sgRNA 结合，而 NTS 结合在 Cas9 蛋白上。最终，HNH 和 RuvC 结构域分别催化 TS 和 NTS 的断裂。

因此，Cas9 识别 PAM 的过程可以看作任意三核苷酸序列（记作 PAM_i）争夺 Cas9 的过程：（1）

$$PAM_i + Cas9 \rightleftharpoons Cas9 \cdot PAM_i \dashrightarrow Product \quad (1)$$

由于在基因组中，任意三核苷酸均可视为 Cas9 识别的潜在目标，这样的三核苷酸数量应当远远大于 Cas9 分子的数量。因此可以认为，在 PAM 识别阶段，Cas9 已被饱和，且被分配到各种三核苷酸（PAM_i）上。而对于各种不同的三核苷酸，如果认为它们

[11] Hu et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. Nature. 2018.
[12] Nishimasu et al. Engineered CRISPR-Cas9 nuclease with expanded targeting space. Science. 2018.
[13] Luan et al. Combined Computational–Experimental Approach to Explore the Molecular Mechanism of SaCas9 with a Broadened DNA Targeting Range. Journal of the American Chemical Society. 2019.

在基因组中的浓度几乎相同，则可以推测 Cas9 结合在某一 PAM_i 上的数量与其结合亲和力相关。通过配分函数可以计算 Cas9 结合在某一 PAM_i 上的比例 P_i：（2）

$$P_i = \frac{e^{-\frac{\Delta G_{PAM_i}}{RT}}}{\sum_j e^{-\frac{\Delta G_{PAM_j}}{RT}}} \quad (2)$$

其中，ΔG_{PAM_i}表示 PAM_i 与 Cas9 的结合自由能，R 是气体常数，T 是温度。由此，即可建立 P_i 与ΔG_{PAM_i}的关系。P_i 应当与 PAM_i 是否能被识别密切相关，若某种 PAM_i 可以占有相当大比例的 Cas9，则可推测它可以被兼容。因此，若可以得到某种 Cas9 与 PAM_i 的结合自由能，则可对其兼容性进行预测。

1．2 基于结构的结合自由能计算

由生物大分子的结构（即原子坐标），可以估算原子间相互作用能量，计算得到生物大分子的自由能。而通过计算 Cas9 和 DNA 结合状态和解离状态自由能之差，则可计算得到结合自由能。我们使用 Rosetta 全原子能量函数（REF）来计算自由能，它同时考虑范德华作用、溶剂化作用、静电作用、氢键、二硫键、二面角等等原子间相互作用的能量，以此估算其自由能。[14]为了由突变体的氨基酸序列来预测其 PAM 序列兼容性，在计算自由能之前，要先对未知结构的突变体进行同源建模。

1.2.1 同源建模

由 SpCas9 的活性状态 cryo-EM 结构（PDB ID:6o0z[15]）作为初始结构，进行同源建模。对于 SpCas9 的突变体（SpCas9-NG 和 xCas9），使用 Rosetta Relax Application 进行同源建模：先对初始结构进行能量最小化，再进行点突变得到 SpCas9-NG 和 xCas9，最后针对突变的残基优化侧链构象。对于 ScCas9，则使用 SWISS-MODEL[16] [17] [18] [19] [20]，以 SpCas9（PDB ID:6o0z）为模板进行同源模建，再利用 Rosetta Relax Application 进行优化。

1.2.2 结合自由能计算

使用 RosettaDNA 计算 64 种 3bp PAM 与 Cas9 的结合能。RosettaDNA 可以对于 PAM 处的 DNA 序列进行突变，并且改变 Cas9 的侧链构象优化 DNA-Cas9 界面处的自由能。对

---

[14]Alford et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. Journal of Chemical Theory and Computation. 2017.

[15]Zhu et al. Cryo-EM structures reveal coordinated domain motions that govern DNA cleavage by Cas9. Nature Structural & Molecular Biology. 2019.

[16] Waterhouse et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018.

[17] Bienert et al. The SWISS-MODEL Repository - new features and functionality. Nucleic Acids Res. 2017.

[18] Guex et al. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. Electrophoresis. 2009.

[19] Benkert et al. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics. 2011.

[20] Bertoni et al. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Scientific Reports. 2017.

于 SpCas9、SpCas9-NG、xCas9 和 ScCas9 均进行了计算，共 256 个结合自由能。

1．3 PAM 特异性预测

由上述结合自由能即可计算结合比例 P_i。将 PAM_i 按照已知实验事实分为两组：可识别的 PAM 和不可识别的 PAM。和预期相符，可识别的 PAM 拥有较大的结合比例，而不可识别的 PAM 结合比例则较小（图 1），两者差异显著。

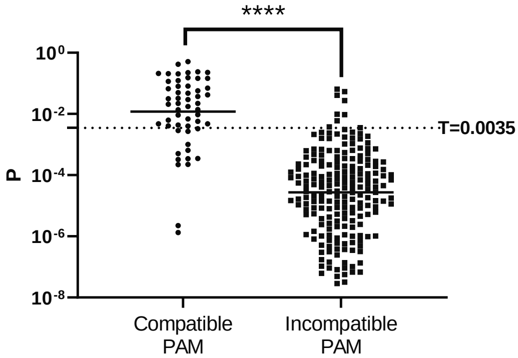
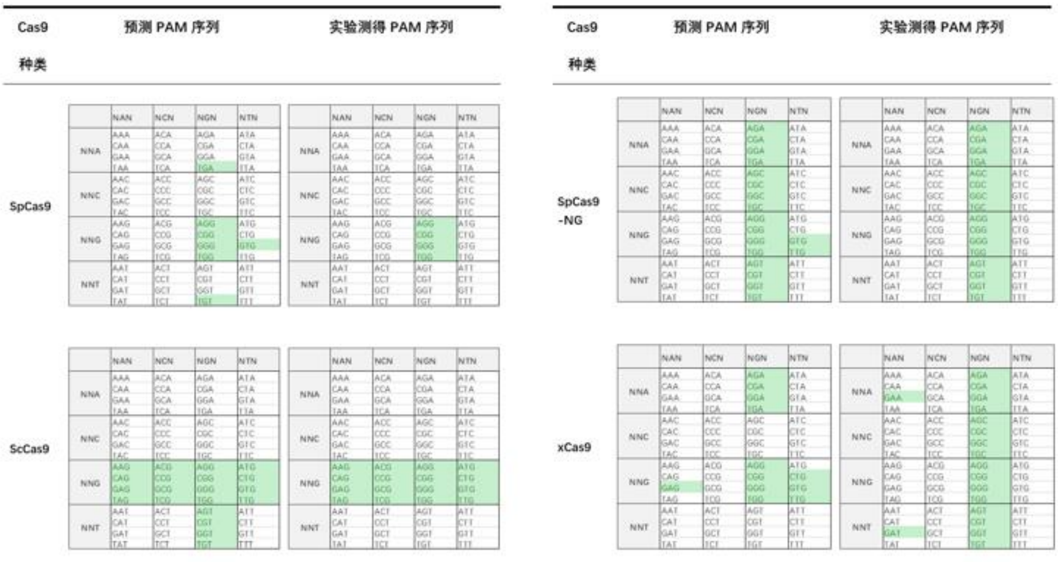


图 1　Cas9 结合比例分布预测

使用 P_i 作为判据，设定某一阈值 T，若 P_i>T，则预测可以识别该 PAM。通过与实验数据比较，可以找到最佳的 T=0.0035，使得对于上述四种 Cas9 同源物，预测错误的概率最小，准确率达到了 91.0%。（表 1）

表 1：预测的 PAM 序列与实验结果比较



GGG=能识别，TTT=不能识别

1．4 SpCas9 新型突变体设计和分析

通过计算机上的随机突变和筛选，我们设计了一个新的 SpCas9 突变体：G1218R，E1219R，R1333V，R1335M，A1320F，P1321Y，A1322R，L1111R，D1135V，由上述模型预测，可以识别 NNA 或 NNG PAM。通过设计引物引入点突变，再进行表达纯化，我们获得了这种 SpCas9 突变体。并且，通过对于靶 DNA 进行点突变，我们获得了含有 5 种不同 PAM 的靶 DNA（TGG、AGG、TAA、TGA、TAG）。然而，活性测试则显示，它对于各种 PAM 均没有活性。（图 2）

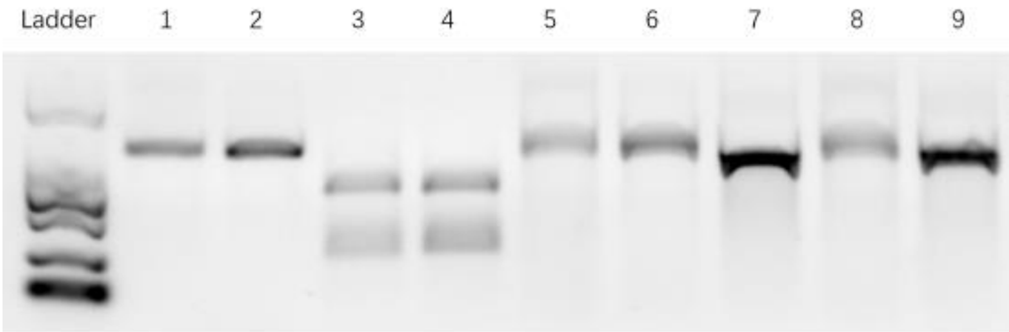


图 2　SpCas9 突变体活性验证

泳道 1、2：靶 DNA。泳道 3、4：SpCas9 和 AGG、TGG PAM 的靶 DNA。泳道 5-9：突变体和 AGG、TGG、TAA、TGA、TAG PAM 的靶 DNA。有活性应显示为双条带，无活性则显示为单条带。

实验结果与模型的预测不符，我们认为可能是如下原因导致的：

1）同源建模结果不准确。由于同源建模只在模板的构象附近很小的构象空间进行构象搜索，它不能处理大幅度的构象改变。如果突变体的骨架采取了另一种折叠方式，同源建模还会按照模板的折叠方式进行建模，很有可能得到错误的模型。

2）模型考虑的参数太少。该模型仅考虑了结合比例 P_i，或者说是结合自由能ΔG_{PAM_i}的相对大小。假设对于某一种 SpCas9 突变体而言，它对于各种不同的 PAM 亲和力都不强，但是结合自由能的分布比较集中，差距不大。该模型会预测它具有很强的 PAM 兼容性，可以兼容大多数 PAM，然而事实上，由于其结合自由能的绝对值过大（亲和力过小），使得它无法和任何 PAM 序列结合。这实际上是由于化学平衡模型的缺陷造成的，它仅能计算催化反应前 Cas9 的平衡状态，而不能考虑催化反应是否发生。

3）由于把所有数据都用作训练集来训练模型（即寻找阈值 T），可能存在过拟合现象，虽然模型与已知数据符合得很好，但是进行预测的能力却很差。

受此启发，我们重新构建了一个 PAM 兼容性预测模型。

2　基于浅层神经网络的 PAM 兼容性预测模型

2．1 输入参数

上述模型仅采用了单个参数 P_i，而本模型中，针对某一种 Cas9 的某一潜在 PAM 序列 PAM_i，使用了 6 个参数：

1) 结合自由能ΔG_{PAM_i}，由 RosettaDNA 得到；

2) 由结合自由能得到的结合比例 P_i；

3) 结合状态总自由能G_{PAM_i}，由 RosettaDNA 得到；

4) 由结合状态总自由能得到的结合比例 PG_i：（3）

$$PG_i = \frac{e^{-\frac{G_{PAM_i}}{RT}}}{\sum_j e^{-\frac{G_{PAM_j}}{RT}}} \quad (3)$$

5) 相对电荷数 Q，与突变体氨基酸序列有关。在 SpCas9 中，Q=0。

6) 相对氨基酸序列长度 L。在 SpCas9 及其突变体中，L=0。在 ScCas9 中，L=83。

由于 P_i 无法反应结合自由能的绝对大小，因此还需考虑结合自由能ΔG_{PAM_i}。考虑结合状态总自由能则是为了反应整体结构的稳定性，减少同源建模错误的风险。考虑相对电荷数是因为静电作用属于长程作用，其效果不能完全反映在 Cas9 和 DNA 的结合界面上。并且 SpCas9-NG 和 ScCas9 都具有更多的正电荷，能更好地吸引带负电的 DNA，它们的 PAM 兼容性较 SpCas9 也更好。考虑氨基酸序列长度则是为了抵消由于 ScCas9 蛋白较大带来的能量差距。

浅层神经网络模型共使用了 9 种 Cas9 作为训练数据：SpCas9、SpCas9-NG、xCas9、ScCas9、VQR、EQR、VRER、D1135E、R1333A/R1335A。（表 2）它们均由上述同源建模的方法生成结构模型，并在此基础上计算各个参数。并且它考虑的潜在识别 PAM（PAM_i）由三核苷酸（NNN，共 64 种组合）改成了四核苷酸（NNNN，共 256 种组合）。总计 256×9=2304 组输入参数。

表 2：训练浅层神经网络模型所采用的九种 Cas9

Cas9 种类	突变位点（相较于 SpCas9）	PAM 序列
SpCas9	/	NGG
SpCas9-NG	R1335V/L1111R/D1135V/G1218R/E1219F/A1322R/T1337R	NG
xCas9	A262T/R324L/S409I/E480K/E543D/M694I/E1219V	NG、GAA、TGT、GAT
ScCas9	SpCas9 同源物	NNG
D1135E	D1135E	NGG
VQR	D1135V/R1335Q/T1337R	NGAN、NGNG
EQR	D1135E/R1335Q/T1337R	NGAG
VRER	D1135V/G1218R/R1335E/T1337R	NGCG

2．2 浅层神经网络的构建、训练与验证

我们使用 MATLAB 的 Neural Net Fitting 工具包来构建和训练浅层神经网络模型。（图 3）上述 6 个参数输入神经网络，经过 6 个神经元的隐层之后，得到一个输出值。将该输出值和阈值 T 相比较，若大于 T，则可预测该输入参数对应的 PAM_i 是可以被识别的，得到阳性预测。

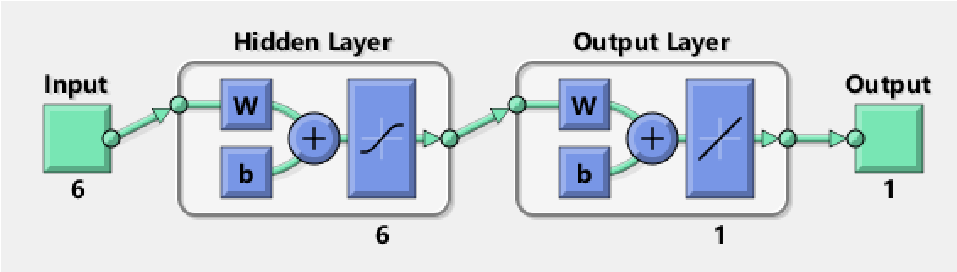


图 3　浅层神经网络模型结构

为了验证模型进行预测的能力，将 2304 组数据随机分为两组：其中 2104 组作为训练集，用于训练模型；200 组作为测试集，在模型的训练中被排除在外，而用于验证模型进行预测的能力。经过训练之后，得到阈值 T=0.553，训练集的准确率为 95.4%。将该模型应用于测试集中，准确率为 93.5%。与之前的模型相比，该模型的准确率有了提升，并且具有一定的预测能力。

2．3 浅层神经网络模型的预测趋势

由该模型的预测趋势可以反推各个参数在模型中的作用。例如，当结合比例 P 比较大、同时结合自由能 ΔG 比较小时，模型才更有可能预测该 PAM_i 可以被识别（阳性）。而在 P 较大，ΔG 也较大时，模型倾向于得到阴性输出结果。（图 4）这就避免了前一个模型中仅考虑 P 而不考虑 ΔG，导致假阳性的现象。

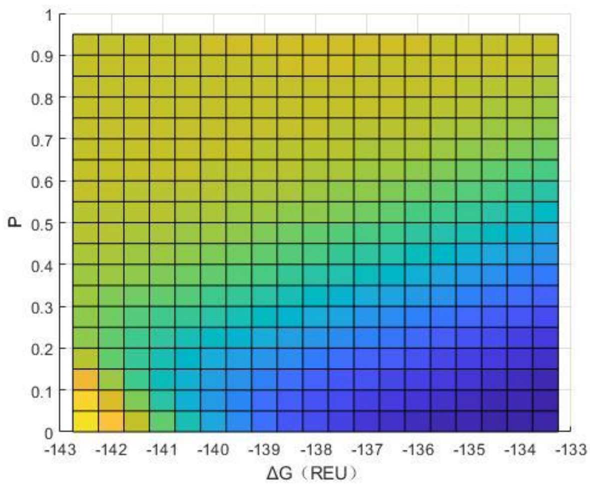


图 4 预测值和结合自由能 ΔG、结合比例 P 的关系

在 SpCas9 的 TGGC PAM 数据组的基础之上，改变 P 和 ΔG 的值进行计算得到。颜色表示预测值大小，预测值越大，越有可能得到阳性的预测结果，即 PAM_i 可被识别。

此外，我们也发现更大的相对电荷数有利于模型做出阳性的预测，这与之前的猜想相符合，即更多的正电荷有利于增强 Cas9 与 DNA 的亲和力。

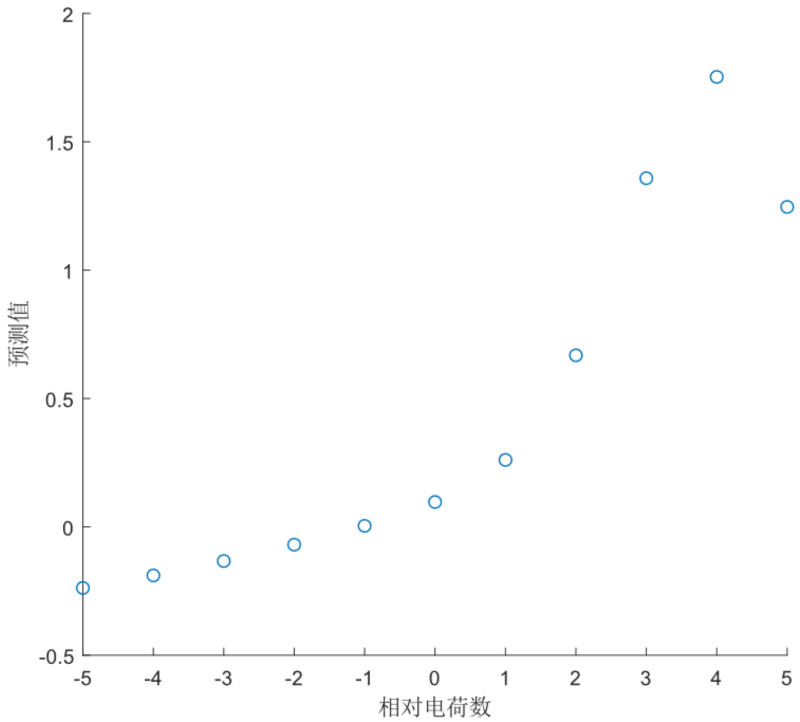


图 5 预测值和相对电荷数的关系

在 SpCas9 的 TGGC PAM 数据组的基础之上，改变相对电荷数的值进行计算得到。由于 P 过大（P=0.46），无论电荷数多少均有很大的预测值，因而此处将 P 调整为 0.1。

总之，该模型在准确率的表现上优于前者，并且新增参数的效果得到了验证，具有更好的 PAM 兼容性预测能力，有一定的应用价值。基于该模型，可以通过计算机随机突变的方法来搜索 PAM 兼容性更好的 SpCas9 突变体。

3 总结

本文探讨了如何结合计算手段来改造SpCas9的PAM序列特异性。通过同源建模、结合自由能计算以及 PAM 识别过程的化学平衡模型，本文构建了一个从 SpCas9 突变体的氨基酸序列预测其 PAM 兼容性的方法，并且尝试利用此方法进行突变体设计。尽管突变体的活性验证失败，我们还是由此发现了模型中的诸多问题。通过引入浅层神经网络模型，我们改良了先前模型中的诸多不足，预测的准确性也有了提高。我们相信，经过不断的改进，这种计算机辅助设计的手段将会成为 Cas9 蛋白设计领域的全新思路。

参考文献

[1] Sorek, R., Lawrence, C. M., & Wiedenheft, B. (2013). CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea. Annual Review of Biochemistry, 82(1), 237-266.

[2] Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J., Charpentier, E., Horvath, P., ... & Koonin, E. V. (2011). Evolution and classification of the CRISPR–Cas systems. Nature Reviews Microbiology, 9(6), 467-477.

[3] Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., Dicarlo, J. E., ... & Church, G. M. (2013). RNA-Guided Human Genome Engineering via Cas9. Science, 339(6121), 823-826.

[4] Gaudelli, N. M., Komor, A. C., Rees, H. A., Packer, M. S., Badran, A. H., Bryson, D. I., & Liu, D. R. (2017). Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature, 551(7681), 464-471.

[5] Anzalone, A. V., Randolph, P. B., Davis, J. R., Sousa, A. A., Koblan, L. W., Levy, J. M., ... & Liu, D. R. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. Nature, 576(7785), 1-1.

[6] Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C., & Doudna, J. A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature, 507(7490), 62-67.

[7] Anders, C., Niewoehner, O., Duerst, A., & Jinek, M. (2014). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature, 513(7519), 569-573.

[8] Jiang, F., Taylor, D. W., Chen, J. S., Kornfeld, J. E., Zhou, K., Thompson, A. J., ... & Doudna, J. A. (2016). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. Science, 351(6275), 867-871.

[9] Chatterjee, P., Jakimo, N., & Jacobson, J. M. (2018). Minimal PAM specificity of a highly similar SpCas9 ortholog. Science Advances, 4(10).

[10] Nishimasu, H., Cong, L., Yan, W. X., Ran, F. A., Zetsche, B., Li, Y., ... & Nureki, O. (2015). Crystal structure of Staphylococcus aureus Cas9. Cell, 162(5), 1113-1126.

[11] Hu, J. H., Miller, S. M., Geurts, M. H., Tang, W., Chen, L., Sun, N., ... & Liu, D. R. (2018). Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. Nature, 556(7699), 57-63.

[12] Nishimasu, H., Shi, X., Ishiguro, S., Gao, L., Hirano, S., Okazaki, S., ... & Nureki, O. (2018). Engineered CRISPR-Cas9 nuclease with expanded targeting space. Science, 361(6408), 1259-1262.

[13] Luan, B., Xu, G., Feng, M., Cong, L., & Zhou, R. (2019). Combined Computational–Experimental Approach to Explore the Molecular Mechanism of SaCas9 with a Broadened DNA Targeting Range. Journal of the American Chemical Society, 141(16), 6545-6552.

[14] Alford, R. F., Leaverfay, A., Jeliazkov, J. R., Omeara, M. J., Dimaio, F., Park, H., ... & Gray, J. J. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. Journal of Chemical Theory and Computation, 13(6), 3031-3048.

[15] Zhu, X., Clarke, R., Puppala, A. K., Chittori, S., Merk, A., Merrill, B. J., ... & Subramaniam, S. (2019). Cryo-EM structures reveal coordinated domain motions that govern DNA cleavage by Cas9. Nature Structural & Molecular Biology, 26(8), 679-685.
[16] Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46(W1), W296-W303 (2018).
[17] Bienert, S., Waterhouse, A., de Beer, T.A.P., Tauriello, G., Studer, G., Bordoli, L., Schwede, T. The SWISS-MODEL Repository - new features and functionality. Nucleic Acids Res. 45, D313-D319 (2017).
[18] Guex, N., Peitsch, M.C., Schwede, T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. Electrophoresis 30, S162-S173 (2009).
[19] Benkert, P., Biasini, M., Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 27, 343-350 (2011).
[20] Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., Schwede, T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Scientific Reports 7 (2017).

致谢：感谢黄强老师和实验室的同学们给我的支持，他们教给我许多实验技能和技巧，组会的讨论也带给我许多灵感，没有他们的帮助我将无法完成这个课题。感谢曦源项目对本课题的资助。