

机器学习方法在 CRISPR/Cas9 系统中的应用

张桂珊, 杨勇, 张灵敏, 戴宪华

中山大学电子与信息工程学院, 广州 510006

摘要: 基于 CRISPR/Cas9 系统介导的第三代基因组定点编辑技术, 已被广泛应用于基因编辑和基因表达调控等研究领域。如何提高该技术对基因组编辑的效率与特异性、最大限度降低脱靶风险一直是该领域的难点。近年来, 机器学习为解决 CRISPR/Cas9 系统所面临的问题提供了新思路, 基于机器学习的 CRISPR/Cas9 系统已逐渐成为研究热点。本文阐述了 CRISPR/Cas9 的作用机理, 总结了现阶段该技术面临的基因组编辑效率低、存在潜在的脱靶效应、前间区序列邻近基序(PAM)限制识别序列等问题, 最后对机器学习应用于优化设计高效向导 RNA (sgRNA) 序列、预测 sgRNA 的活性、脱靶效应评估、基因敲除、高通量功能基因筛选等领域的研究现状与发展前景进行了展望, 以期对基因组编辑领域的研究提供参考。

关键词: CRISPR/Cas9; 机器学习; sgRNA; 脱靶效应; 基因敲除

Application of machine learning in the CRISPR/Cas9 system

Guishan Zhang, Yong Yang, Lingmin Zhang, Xianhua Dai

School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China

Abstract: The third generation of the CRISPR/Cas9-mediated genome fixed-point editing technology has been widely used in the field of gene editing and gene expression regulation. How to improve the on-target efficiency and specificity of this system, as well as reduce its off-target effects are always the bottleneck in its development. Machine learning provides novel methods to the problems of the CRISPR/Cas9 system, and CRISPR/Cas9-based machine learning has recently become a very hot research topic. In this review, we firstly outline the mechanism of the CRISPR/Cas9 system. Subsequently, we elaborate the current issues of CRISPR/Cas9, including low efficiency and potential off-target effects, and sequence-recognizing limitation from protospacer adjacent motif (PAM). Finally, we summarize the applications of methods within the machine learning framework for optimizing the CRISPR/Cas9 system, such as optimized single-guide RNA (sgRNA) design, CRISPR/Cas9 cleavage efficiency prediction, off-target effects evaluation, gene knock-out as well as high-throughput functional genetic screening and prospects for development.

Keywords: CRISPR/Cas9; machine learning; sgRNA; off-target effect; gene knock-out

收稿日期: 2018-05-15; 修回日期: 2018-07-19

基金项目: 国家自然科学基金项目(编号: 61872396)资助[Supported by National Natural Science Foundation of China (No.61872396)]

作者简介: 张桂珊, 博士研究生, 研究方向: 生物信息学。E-mail: zhanggsh7@mail2.sysu.edu.cn

通讯作者: 戴宪华, 博士, 教授, 研究方向: 生物信息学。E-mail: issdxx@mail.sysu.edu.cn

DOI: 10.16288/j.ycz.18-135

网络出版时间: 2018/7/30 15:48:00

URI: <http://kns.cnki.net/kcms/detail/11.1913.R.20180730.1548.004.html>

CRISPR/Cas9 (clustered regularly interspaced short palindromic repeat/CRISPR-associated protein 9) 系统介导的基因组编辑技术^[1-3]是继锌指核酸酶(zinc-finger nucleases, ZFNs)、类转录激活因子效应物核酸酶(transcription activator-like effector nuclease, TALENs)后出现的第三代“基因组定点编辑技术”,可对特定位置上的DNA序列进行编辑与修改。近年来,CRISPR/Cas9技术凭借成本低廉、容易操作等优点迅速成为基因工程领域的关注热点。目前,CRISPR/Cas9技术主要应用于基因敲除、基因敲入、DNA大片段删除、转录调控、基因检测、基因标记。但是,该技术仍存在许多科学问题有待研究。比如,CRISPR/Cas9是一种单链酶,其自身具有不稳定性,容易引发突变导致脱靶效应^[4, 5]。而且,不同靶点Cas9内切酶切割效率存在明显的差异^[6-10],靶点位于核小体核心区域,Cas9的活性受到抑制;靶点位于DNA邻接区,核小体结构不影响Cas9的活性^[7]。此外,CRISPR/Cas9可能在靶点远处删除大规模DNA^[11]。因此,克服脱靶效应和提高基因组编辑效率成为研究人员亟待解决的问题。

机器学习(machine learning, ML)是人工智能(artificial intelligence, AI)的核心,主要研究如何通过计算的方法,利用经验改善系统的性能^[12]。机器学习能有效地分析经验数据,为生物信息学提供重要的技术支撑。基于机器学习预测CRISPR/Cas9脱靶效应的主要思路,是利用先验知识(priori knowledge)学习引起脱靶效应的sgRNA与目标基因组序列碱基错配数目的统计学规律,进而评估sgRNA的脱靶效应^[13]。此外,机器学习在CRISPR/Cas9系统优化设计高效sgRNA序列^[14-16]、预测sgRNA活性^[17]、设计CRISPR干扰(CRISPR interference, CRISPRi)/CRISPR激活(CRISPR activation, CRISPRa)高效率的sgRNA^[8]、设计全基因组CRISPR/Cas9基因敲除文库^[18]、预测CRISPR重复序列的方向^[19]、鉴定必需基因^[18, 20-24]等方面有着日渐广泛的应用。本文在阐述CRISPR/Cas9系统作用机制的基础上,对现阶段CRISPR/Cas9研究领域所存在的科学问题进行讨论与分析,然后对机器学习应用于CRISPR/Cas9系统设计高效sgRNA、脱靶位点预测、打靶活性评估、基因敲除、高通量功能基因筛选等展开综述,以期对相关领域的研究提供参考。

1 CRISPR/Cas9系统的组成及其作用机制

CRISPR序列最早发现于细菌和古细菌中,几乎所有的古细菌和40%的细菌都具有此类序列^[25, 26]。CRISPR序列中含有大量的重复序列和间隔序列(proto-spacer),其间隔序列长度大致相同,并具有特异性^[27]。CRISPR/Cas9是细菌和古细菌在长期演化过程中形成的一种适应性免疫防御系统,用于对抗入侵的病毒及外源DNA。CRISPR/Cas9系统通过将入侵噬菌体和质粒DNA片段整合到CRISPR中,指导相应的CRISPR RNA(crRNA)降解同源序列,同时该系统具有免疫记忆能力^[28, 29]。

CRISPR/Cas9系统由CRISPR序列元件和Cas9核酸酶组成。Cas9核酸酶在crRNA (CRISPR RNA)和反式激活crRNA (trans-activating crRNA, tracrRNA)的指导下,在具有前间区序列邻近基序(proto-spacer adjacent motif, PAM)的DNA双链靶点(PAM上游3个碱基处)进行靶向双链切割,形成钝末端DNA双链断裂(double strand breaks, DSBs)^[25, 30, 31]。其作用过程大致如图1所示:首先,Cas9蛋白复合物与crRNA和tracrRNA组成功能复合物,在crRNA引导下靶向识别并结合具有PAM位点相匹配的DNA靶序列。然后,Cas9依靠其自身的两个核酸内切酶结构域发挥内切酶的作用,在PAM上游约3个碱基位点处进行切割。在切割过程中,Cas9核酸酶的HNH (His-Asn-His)结构域切割模板链,RuvC (核糖核酸酶H/整合酶超家族的成员)结构域切割非模板链,进而导致目标DNA双链断裂^[32]。最后,引发细胞启动自动修复机制。通过非同源末端连接(nonhomologous end-joining, NHEJ),细胞引入插入/缺失(indel)碱基引起靶点位置基因的突变;通过同源重组修复(homology-directed repair, HDR),细胞利用外源DNA提供的“供体模板”与突变靶点重组,实现对基因组的DNA定点编辑^[33-35]。

2 CRISPR/Cas9基因组编辑效率与特异性

2.1 PAM序列对CRISPR/Cas9系统基因组编辑效率与特异性的影响

CRISPR/Cas9切割位点的编辑效率与特异性依

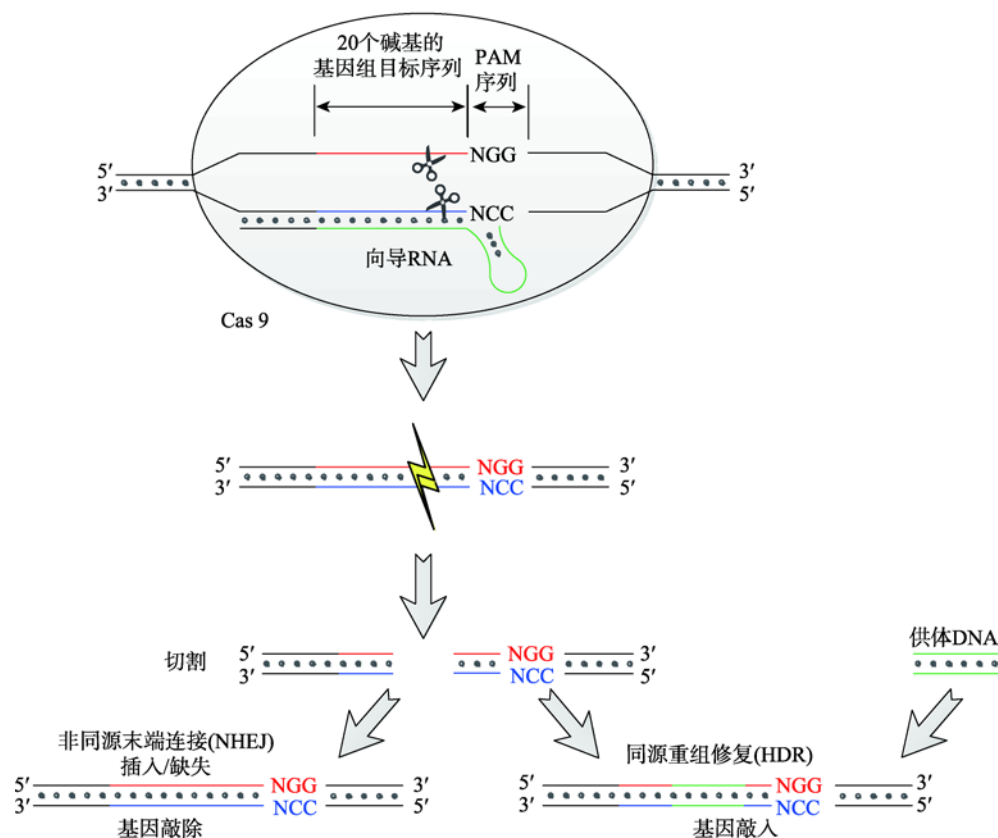


图 1 CRISPR/Cas9 对靶基因的切割示意图

Fig. 1 Schema of CRISPR/Cas9-mediated cleavage of target genes

赖于 PAM 序列^[36]。在 CRISPR/Cas9 系统中, PAM 序列是位于靶点 DNA 链 3'端的 3 个碱基(通常为 NGG, N = A, C, G 或 T), PAM 在该系统中发挥着关键作用。研究表明, PAM 序列有助于 Cas9 结合与切割目标 DNA 序列^[37], Cas9 利用 PAM 快速识别潜在的靶点。若目标 DNA 序列 3'端不存在 PAM, 即使目标序列与 sgRNA 序列完全匹配, Cas9 也无法切割该序列靶点^[37]。CRISPR/Cas9 系统“经典 PAM 序列”(canonical PAM)为 NGG; “非经典 PAM 序列”(non-canonical PAM), 如 NAG、NCG 和 NGA 也能够激发 Cas9 的活性^[17, 38~40]。PAM 类型影响 CRISPR/Cas9 的目标 DNA 序列编辑效率^[38]。Zhang 等^[38]发现, PAM 影响的基因组编辑效率大小依次为 NGG > NGA > NAG, Doench 等^[22]发现 NGG (97%) > NAG (26%) > NCG (11%) > NGA (7%)。

CRISPR/Cas9 系统近 PAM 端 1~12 位碱基的序列称为“种子序列”(seed sequence)^[5, 36, 39, 41], 近 PAM 端 1~5 位碱基序列是更精确的种子区域, 该子区域

被称为“核心区域”(core region)^[42, 43]。与远 PAM 端碱基相比, 近 PAM 端的碱基对 CRISPR/Cas9 的特异性影响更大。Jinek 等^[36]发现 Cas9 对种子序列区域 sgRNA 与目的 DNA 碱基错配的耐受能力较差, 而对非核心区的碱基错配具有较好耐受能力。Hsu 等^[39]研究发现, 在近 PAM 端 (PAM 序列 5'端第 1 至 10~12 位碱基), sgRNA 与目标 DNA 序列碱基配对数目影响 Cas9 的特异性。近 PAM 端 1~12 位碱基区域, sgRNA 与目标 DNA 序列碱基错配导致靶点切割效率降低(消失), 远 PAM 端序列错配也可能剪切该位点^[1]。Zhang 等^[39]发现, 紧邻 PAM 上游 8~14 个碱基序列是影响 CRISPR/Cas9 特异性的关键因素。近 PAM 端两个碱基或 3 个碱基突变, 无论是相邻的还是分散的, 对 CRISPR/Cas9 特异性影响较大。而且, 错配碱基数目大于 3 时, CRISPR/Cas9 切割基本消失。Mali 等^[7]发现, 在非种子序列区域, Cas9 可以容忍 sgRNA 与目标 DNA 序列存在 1~3 个碱基错配; 而在种子序列区域, 错配两个碱基导致 Cas9 活

性降低。此外, Cas9 对 sgRNA 与目标 DNA 序列错配碱基数目的耐受能力与反应条件有关。当 sgRNA 和 Cas9 浓度较高时, sgRNA 与目的 DNA 错配碱基数目不能大于 5 个碱基^[4]。PAM 如何影响该系统基因组编辑效率的作用机制将有待进一步研究。

2.2 sgRNA 对 CRISPR/Cas9 系统基因组编辑效率与特异性的影响

CRISPR/Cas9 系统靶点识别特异性主要依赖 sgRNA, 因此, 设计与选择 sgRNA 是 CRISPR/Cas9 技术成功的关键。sgRNA 的编辑效率与特异性受诸多因素的影响: (1) sgRNA 上 20 个碱基的向导序列 (guide RNA, gRNA) 对 Cas9 靶向性和特异性有着重要的作用^[22, 24], 近 PAM 端 gRNA 与目的 DNA 碱基配对数目决定 Cas9 的特异性。(2) 考虑微同源特征 (microhomology feature) 能够提高 sgRNA 的活性^[44]。Doench 等^[45]分析了影响 sgRNA 效率活性的序列特征, 发现在第 20 号位置的 sgRNA 偏好 C 而排斥 G, 在 sgRNA 的中间区域, 存在 A 的 sgRNA 活性较高。此外, 富含 G 且仅含少数 A 的 sgRNA 稳定性和活性较高^[46]。然而, 过高或过低的 GC 含量均导致 sgRNA 的编辑活性降低^[45, 47]。(3) sgRNA 序列长度影响 CRISPR/Cas9 的剪切特异性。适当改变 sgRNA 的长度, 在识别序列前增加两个 G^[48]或截断 5' 端 2~3 个碱基^[48, 49], 能够减少脱靶效应^[22, 39]。(4) 优化 sgRNA 的结构可以提高 CRISPR/Cas9 的编辑效率。利用 double-nickase 策略, 突变 Cas9 (D10A), 设计成对 sgRNA 从而提高 CRISPR/Cas9 的编辑效率。而且, 两条 sgRNA 的剪切位点之间的距离越近, 编辑效率越高^[50, 51]。(5) 结合表观遗传修饰特征可以提高 sgRNA 的编辑效率。如蛋白质紊乱状态 (protein disorder status) 影响 sgRNA 与目的 DNA 相结合^[24], 研究表明, 结合染色质的易接近性 (脱氧核糖核酸酶超敏位点, DNase I hypersensitive sites) 可以提高 gRNA 的活性^[14]。

目前, 利用计算方法设计编辑效率高的 sgRNA 主要考虑 sgRNA 序列特征及其二级结构特征, 将所有可能的 sgRNA 按分数排序, 选择具有较高切割效率的 sgRNA^[52]。自 2013 年以来, 已有多款在线或单机版 sgRNA 设计和脱靶效应评估软件, 包括

CRISPR DESIGN^[39]、sgRNAcas9^[53]、Protospacer^[54]、Cas9 Design^[55]、CCTop^[27]和 CFD^[22]等。

2.3 CRISPR/Cas9 系统存在潜在的脱靶效应

CRISPR/Cas9 系统靶向生物基因组存在潜在的脱靶效应^[17, 39, 48, 49, 56, 57]。Cas9 核酸酶对 sgRNA 与目标 DNA 序列碱基匹配具有一定的容错能力^[39]。sgRNA 除了正常切割靶点 DNA 双链, 也可能与靶点同源性较高的非靶点 DNA 序列局部匹配 (partial match), 激活 Cas9 切割非目标序列, 产生脱靶效应^[39]。Hsu 等^[39]发现脱靶效应主要取决于 sgRNA 与靶 DNA 序列错配碱基的数目、错配碱基所在位置、碱基错配类型 (转换、颠换) 等。存在非经典 PAM 序列且 sgRNA 与靶 DNA 序列碱基错配数目较大时, CRISPR/Cas9 脱靶可能发生在非目的位点, 甚至脱靶位点的编辑效率高于目标靶点^[52]。

目前, 已有诸多研究策略旨在降低 CRISPR/Cas9 系统的脱靶效应, 主要包括: (1) 选择合适的靶点。合理地避开 GC 含量较高的靶点可以降低脱靶效应^[1]。(2) 适当改变 sgRNA 的长度。(3) 控制 sgRNA 与靶 DNA 序列碱基错配数目。(4) 合理控制 Cas9 与 sgRNA 表达的剂量与持续时间。通过滴定控制 Cas9 和 sgRNA 的表达剂量, 可以有效地降低脱靶效应^[39]。当 Cas9 与 sgRNA 的复合物含量较低时, 脱靶效应较低^[5]。(5) 利用全基因组无偏鉴定方法分析脱靶效应, 包括基于寡核苷酸整合成双链断裂 (GUIDE-Seq)^[56, 58], 高通量全基因组易位测序 (HTGTS)^[59]等。

3 机器学习在 CRISPR/Cas9 系统的应用

自 2014 年以来, 机器学习已逐步应用于优化设计高效 sgRNA 序列、预测 sgRNA 的活性、脱靶位点预测、基因敲除、高通量功能基因筛选等。运用机器学习优化 CRISPR/Cas9 系统主要分 6 个步骤 (图 2): 第一, 整合不同实验平台、数据库的数据; 第二, 数据预处理 (数据清洗), 剔除冗余信息; 第三, 构造特征, 由 sgRNA 与靶序列碱基配对情况、sgRNA 序列本身的特征与二级结构特征、染色质状态等构造特征集; 第四, 特征选择, 运用特征选择算法选取与目标相关的特征; 第五, 训练模型, 通过交叉

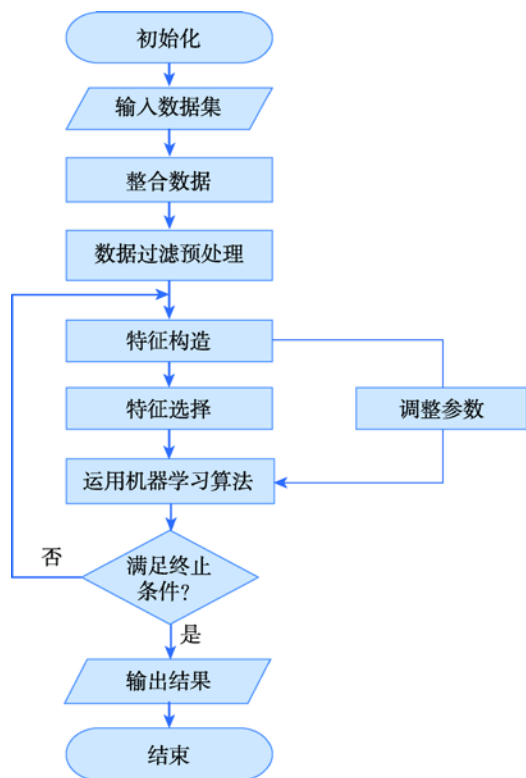


图 2 机器学习优化 CRISPR/Cas9 系统流程图
Fig. 2 Schematic flow of the machine learning method for optimizing the CRISPR/Cas9 system

验证判断模型是否存在过拟合(overfitting)/欠拟合(underfitting), 调整参数, 进一步优化模型; 第六, 模型分析, 通过实验测试进行误差分析, 评估所得模型的有效性、泛化性。

3.1 数据整合

自 2014 年以来, 可用于研究机器学习在 CRISPR/Cas9 系统中的应用的开源数据集与在线资源逐年增加, Hart 等对此进行了详细的汇总。例如, FC_RES 数据集包含 17 种基因, 共含有 4380 条 sgRNA 序列, 包括 6 种大鼠基因(Cd5, Cd28, H2-K, Cd45, Thy1, Cd43)、11 种人类基因(CD13, CD15, CD33, CCDC101, MED12, TADA2B, TADA1, HPRT, CUL3, NF1, NF2)。FC 数据集包含人类和大鼠 9 种基因共 1841 条 sgRNA 序列, RES 数据集包含 8 种基因, 共 2549 条 sgRNA 序列^[60]。机器学习已应用 FC_RES 数据集预测 sgRNA 的脱靶效应^[13]和打靶活性^[60], 优化了 sgRNA 编辑效率^[24], 且进一步结合表达策略预测 sgRNA 的编辑效率^[61]等。表 1 总结了常用的机器学习方法应用于 CRISPR/Cas9 系统研究所采用的数据集。

表 1 优化 CRISPR/Cas9 系统的机器学习方法常用数据集

Table 1 Major datasets of machine learning methods for optimizing the CRISPR/Cas9 system

工具名	年份	Cas 类型	数据集/数据集来源	数据集 URL	参考文献
Elevation	2018	Cas9	FC, "Schönig", "Concordet", "Eschstruth", "Shkumatava" U2OS, HEK293	https://www.ncbi.nlm.nih.gov/sra?term=SRP117146%5BAccession%5Dhttps://github.com/maximilianh/crisporWebsite	[13]
DeepCpf1	2018	Cpf1	HT 1-2, HT 2, HT 3, HEK-lenti, HEK-plasmid, HCT-plasmid HEK293T cells	http://www.rgenome.net/cpf1-database	[14]
CRISTA	2017	Cas9	GUIDE-seq data, BLESS data, HTGTS data HEK293, U2OS	http://crista.tau.ac.il/	[52]
CRISPRpred	2017	Cas9	FC, RES	http://research.microsoft.com/en-us/projects/azimuth/	[60]
sgRNA Designer (Rule Set 2)	2016	Cas9	FC, RES A375, HL60, KBM7, mouse ESC JM8	https://www.nature.com/articles/nbt.3437#supplementary-information	[22]
predictSGRNA	2017	Cas9	ribosomal genes, non-ribosomal genes, essential genes HL-60, KBM-7, mouse ESC JM8	http://genome.cshlp.org/content/25/8/1147/suppl/DC1 http://www.sciencemag.org/content/343/6166/80/suppl/DC1 http://www.nature.com/nbt/journal/v32/n3/full/nbt.2800.html#supplementary-information	[23]
Big Papi	2017	Cas9	A375, 293T, MOLM13	https://github.com/mhegde	[16]
—	2017	Cas9	FC, RES, UniRef100	http://research.microsoft.com/en-us/projects/azimuth	[24]
sgRNA Scorer 2.0	2017	Cas9 Cpf1	293T	https://pubs.acs.org/doi/abs/10.1021/acs.synbio.6b00343	[62]

续表

工具名	年份	Cas 类型	数据集/数据集来源	数据集 URL	参考文献
CRISPR-DO	2016	Cas9	full human (GRCh37/hg19, GRCh38/hg38) mouse (NCBI37/mm9 GRCm38/mm10) zebrafish (danRer7), fly (dm6), worm (ce10) HL60, 293T, KBM7	https://www.ncbi.nlm.nih.gov/pubmed/?term=CRISPR-DO+for+genome-wide+CRISPR+design+and+optimization	[63]
CRISPR multitargeter	2015	Cas9	BioMart, zebrafish ohnologs	https://github.com/SergeyPry/CRISPR_MultiTargeter	[64]
CRISPRscan	2015	Cas9	One-cell-stage zebrafish embryos, germ	https://www.nature.com/articles/nmeth.3543	[46]
WU-CRISPR	2015	Cas9	FC 293T, K562, A549, HepG2, SKNAS, U2OS, PGP1-iPS, HEK293	https://www.ncbi.nlm.nih.gov/sra/	[6]
CRISPR (SSC)	2015	Cas9	HL60, KBM7, ABL1, BCR, 293T, LNCaP-abl, mouse ESC JM8 A promoter-level mammalian expression atlas; Determinants of nucleosome organization in primary human cells; An integrated encyclopedia of DNA elements in the human genome	http://fantom.gsc.riken.jp/5/datafiles/phase1.3/extra/TSS_classifier/ https://www.encodeproject.org/files/ENCFF00VNN https://www.encodeproject.org/files/ENCFF000TLU	[65]
CRISPRko	2014	Cas9	FC A375, EL4, AML, MOLM13, NB4, TF1	https://www.nature.com/articles/nbt.3026#supplementary-information	[45]
—	2014	Cas9	HL60, 293T, KBM7, DH5 α	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3972032/#SD1	[47]
SgRNA Scorer 1.0	2015	Cas9	DNase-seq (GSM1008573), H3K4-trimethylation (GSM945288) 293T, K562, A549, HepG2, SKNAS, U2OS, PGP1-iPS, HEK293	http://arep.med.harvard.edu/CasFinder/	[66]
CRoatan	2017	Cas9	FC A375, K562	http://dx.doi.org/10.1016/j.molcel.2017.06.030	[61]
TKOv3	2017	Cas9	essential genes, nonessential genes CEG2, KBM7, HL60, RPE1, DLD1, GBM, HAP1, HCT116, RPE1dTP53	http://tko.ccbbr.utoronto.ca/	[18]
BAGEL	2016	Cas9	essential genes, nonessential genes GBM, HCT116, HeLa, RPE1	http://tko.ccbbr.utoronto.ca/	[20]
CRISPRiaDesign	2016	Cas9	FANTOM Consortium, ENCODE; Consortium (accession no. ENCF000VNN); ENCODE Consortium (accession no. ENCF000TLU); K562, HEK293T; A promoter-level mammalian expression atlas; Determinants of nucleosome organization in primary human cells	http://fantom.gsc.riken.jp/5/datafiles/phase1.3/extra/TSS_classifier/ https://www.encodeproject.org/files/ENCFF00VNN https://www.encodeproject.org/files/ENCFF000TLU	[8]
CRISPRstrand	2014	Cas	REPEATS _{Lange} ; REPEATS _{Kunin} ; REPEATS _{Shah}	http://www.ncbi.nlm.nih.gov/	[19]
H1/H2 library	2018	Cas9	K562, Raji, Jiyoye, KBM7	https://doi.org/10.1093/bioinformatics/bty450	[67]

—: 代表该文献未定义该算法工具的名称。

3.2 数据预处理

由于计算机只能处理数字信号, 利用计算的方法分析 CRISPR/Cas9 系统, 需要将 DNA/RNA 符号序列数值化。选择合适的映射机制, 在保留生物学信息的前提下, 将 DNA/RNA 符号序列转换成数值序列。由于不同实验平台数据的格式不一致, 故应选择适当的变换方法进行数据预处理分析。

3.2.1 数据变换

当数据存在偏移时, 先对原始数据进行变换(如对数变换、Box-Cox 变换^[68]), 使该数据服从正态分布。Abadi 等^[52]将 Kleinsteiver^[5], Frock^[48], Ran^[69]和 Slaymaker^[70]的数据统一整理成 Tsai^[71]的格式。Abadi 等^[52]做切割效率预测分析, 为了强化学习过程, 仅保留序列比对分数大于 14.75 的位点(95%的切割位点序列匹

配个数均值为 16.7)。Doench 等^[22]在评估基因敲除时, 将 sgRNA 分数进行尺度变换成[0, 1], 1 表示成功敲除基因。

3.2.2 单字母编码法

单字母编码法(one-letter code)^[24]对 4 种核苷酸(A、C、G、T)进行编码, 将长度为 N 的 sgRNA 序列编码成一个 4 维向量, 依据碱基存在性将向量的每一个分量设置为 0 或 1。如 sgRNA 中“G”编码为“0010”。单字母编码法将第 $[i, i+1]$ ($i=1, 2, \dots, N-1$) 个所有可能的成对核苷酸(AA, AC, ..., TG, TT)分别编码成一个 16 维的向量。如“CG”编码成“0000001000000000”。长度为 30 个碱基的 sgRNA 编码成 584 ($30 \times 4 + 29 \times 16$) 个单核苷酸和双核苷酸对序列。

3.2.3 独热编码法

独热编码(one-hot coding)通过有效地增加额外的列, 分别用 0 和 1 表示分类值的有或无。Doench 等^[22]采用独热编码, 对于一阶特征, 将长度为 30 bp 的 sgRNA 第一个位置可能出现的 A/C/G/T 转换成 4 个二进制变量, 分别代表一种可能的核苷酸。二阶特征(所有相邻的两个核苷酸作为一个特征: AA/AT/AG 等)共 $4 \times 4 = 16$ 个。独热编码将每个二阶特征变编码成一个 16 位二进制变量。

3.3 特征构造

机器学习应用于 CRISPR/Cas9 系统 sgRNA 设计、sgRNA 活性预测、脱靶效应评估、基因敲除、高通量功能基因筛选等, 特征选择的目标大致如下: 第一, 提高算法预测的准确性; 第二, 构造计算低复杂度的机器学习预测模型; 第三, 更好地理解解析模型。运用机器学习分析 CRISPR/Cas9 系统主要考虑以下特征: (1)序列的特征, 如序列本身的特征(序列中 GC 含量、ACGT 各碱基数)、位置依赖的特征(sgRNA 中与位置有关的单核苷酸、二核苷酸、3 个连续的核苷酸), 以及 sgRNA 与目标 DNA 序列碱基错配类型^[6, 8, 13, 14, 16, 19, 22-24, 42, 43, 45-47, 52, 60, 61, 63-66]。(2)热力学特征, 包括局部配对概率、sgRNA 的最小自由能量(minimum free energy)^[23, 52, 60, 64], DNA 焓(DNA enthalpy)^[52]。DNA 焓与染色质的状态有关,

反映基因组位点或附近双螺旋的结合亲和力。DNA 焓在预测 Cas9 效率方面可能有着重要的作用。研究表明, DNA 焓影响转录因子与其他 DNA 结合蛋白的结合^[52]。但是, 其对 Cas9 蛋白的亲和力的贡献还不明确。(3)核染色质的状态, 如超敏位点^[13, 14, 43, 52, 64]。(4)氨基酸的切割位置(amino acid cut position)和肽百分比(percent peptide)^[24, 60]。(5)核小体占位(nucleosome occupancy)^[7, 8, 33, 34, 72]。(6)微同源特征(microhomology feature)^[22]。(7)蛋白质紊乱状态^[24]。(8)靶位点与 PAM 序列的距离^[62, 64]。(9)切割位点核苷酸的偏好性^[42, 63]。(10)外显子类型^[47]。(11) sgRNA 靶向必需基因与非必需基因的概率分布^[18, 20]。(12)序列折叠成二级结构的倾向性^[5]。常用优化 CRISPR/Cas9 系统的机器学习算法及其在该系统的应用与所构建的特征见表 2。

3.4 特征选择

利用机器学习优化 CRISPR/Cas9 系统, 特征集往往较大, 特征之间可能存在相关性。训练特征过多导致训练模型所需时间较长, 容易引起“维度灾难”, 导致模型复杂, 泛化性差^[73]。机器学习预测模型与特征选择息息相关, 选择不同的特征训练将得到不同的模型^[74]。

特征选择(feature selection)^[75]旨在从特征集中选取一个代表全局特征的子集。特征选择主要分为 4 个部分(图 3): (1)产生过程; (2)评价函数; (3)停止准则; (4)验证过程。首先, 从特征集中选取一个特征子集。利用评价函数对该子集进行计分并与停止准则进行比较, 若满足停止条件, 程序结束; 否则继续产生下一组特征子集, 重复上述步骤。最后, 验证特征子集的有效性。评价函数主要包括三种方法: 筛选器(filter)、封装器(wrapper)及嵌入法(embedded)。筛选器对每一维的特征赋予权重, 依据权重排序, 如信息增益、相关系数。封装器将选择子集视为一个搜索寻优问题, 生成不同的特征组合分别进行评价与比较。嵌入法在模型给定的情况下, 学习提高模型准确性的属性。正则化对权重进行约束, 岭回归(ridge regression)在线性回归中加入了正则项。停止准则通常是一个与评价函数有关的阈值, 当评价函数值达到该阈值则停止搜索。验证过程主

表 2 常用优化 CRISPR/Cas9 系统的机器学习方法总结

Table 2 Summary of optimizing CRISPR/Cas9 system methods based on machine learning

工具名	年份	机器学习方法	机器学习方法在设计该工具的应用	特征构造与选择	参考文献
Elevation	2018	朴素贝叶斯; 梯度提升回归树; L1 正则线性回归	预测 sgRNA 的脱靶效应	sgRNA 与目的 DNA 碱基错配位置; sgRNA 与目的 DNA 碱基错配类型; 与位置相关的 sgRNA 与目的 DNA; 碱基错配类型; 突变类型(转换、颠换); 染色质的易接近性	[13]
DeepCpf1	2018	卷积神经网络	预测 CRISPR-Cpf1 sgRNA 的编辑效率	sgRNA 序列特征; 染色质的易接近性	[14]
CRISTA	2017	回归模型; 随机森林	评估 sgRNA 切割 基因组位点的倾向性	PAM 序列类型, sgRNA 序列核苷酸 组成, GC 含量; 染色质的结构, 编码区基因表达水平; sgRNA 的二级结构、热力学特征; sgRNA 与目标 DNA 序列错配; DNA 凸起/RNA 凸起的数量	[52]
CRISPRpred	2017	支持向量机(SVM); 逻辑回归; 随机森林	预测 sgRNA 的 在靶活性	单核苷酸、双核苷酸、三个连续的 核苷酸在 sgRNA 中的位置; 最小自由能、局部配对概率、sgRNA 的热量; sgRNA 序列 GC 含量、AT 含量、A/C/G/T 数; 氨基酸的切割位置, 肽所占百分比	[60]
sgRNA Designer (Rule Set 2)	2016	线性回归; L1/L2 正则逻辑回归; 支持向量机; 随机森林; 梯度上升回归树	预测 sgRNA 的 在靶活性	二核苷酸特征; 与位置有关的单核苷酸和双核苷酸; sgRNA 中 GC 含量; 位置独立的核苷酸数; sgRNA 靶点在基因中的位置; 微同源特征	[22]
predictSGRNA	2017	逻辑回归; 随机森林	设计高编辑效率 sgRNA	位置依赖的单核苷酸; 位置独立的单核苷酸; 单核苷酸与二核苷酸的频率; sgRNA 与目的 DNA 比对得分; 热力学特征及二级结构、理化性质; 由 PseKNC 模型生成伪 k-元组核苷酸特征	[23]
Big Papi	2017	梯度上升回归树	优化设计 sgRNA 文库	位置独立的单核苷酸、二核苷酸; 位置依赖的单核苷酸、二核苷酸; 热力学特性(解链温度); 3'PAM 序列最接近的胸腺嘧啶(T)	[16]
—	2017	最小冗余最大相关性; 优化 支持向量机	优化 sgRNA 编辑效率	单个、成对的核苷酸(SNTs, PNTs); sgRNA 与目的 DNA 序列保守性; 氨基酸切割位置, 编码肽的氨基酸组成; 靶蛋白序列的紊乱状态	[24]
sgRNA Scorer 2.0	2017	支持向量机	预测 sgRNA 编辑效率	靶点与 PAM 序列的距离	[62]
CRISPR-DO	2016	LASSO 回归; 弹性网络线性回归	预测 sgRNA 的 编辑效率	sgRNA 序列特征; 切割位点处胞嘧啶的偏好性	[63]
CRISPR multitargeter	2015	逻辑回归	预测 sgRNA 的 编辑效率	GC 百分比, 与位置有关的单核苷酸、 相邻的二核苷酸; sgRNA 中 G, C 的含量, G/C 比值; 局部染色质结构	[64]

续表

工具名	年份	机器学习方法	机器学习方法在设计该工具的应用	特征构造与选择	参考文献
CRISPRscan	2015	逻辑回归	预测 sgRNA 的编辑效率	位置依赖的单核苷酸、二核苷酸；GC 含量	[46]
WU-CRISPR	2015	支持向量机	预测 sgRNA 的编辑效率与编辑特异性	位置独立的单核苷酸、二核苷酸；位置依赖的单核苷酸、二核苷酸；RNA 的二级结构(折叠自由能、核苷酸的易接近性)；位置依赖的 sgRNA 核苷酸的易接近性；sgRNA 重复碱基的分布	[6]
CRISPR (SSC)	2015	LASSO 回归	预测全基因组功能基因筛选 sgRNA 的编辑效率	sgRNA 序列特征；切割位点处胞嘧啶的偏好性	[65]
CRISPRko	2014	支持向量机；逻辑回归	预测 sgRNA 的编辑效率	与位置有关的单核苷酸、相邻的二核苷酸；sgRNA 中 G, C 的含量, G/C 比值；局部染色质结构	[45]
—	2014	支持向量机	预测 sgRNA 的编辑效率	序列特征, GC 含量；sgRNA 链, 外显子类型	[47]
SgRNA Scorer 1.0	2015	支持向量机	分析不同活性(高/低)位置依赖的序列特征 sgRNA 之间的关系研究 sgRNA 特异性与活性的关系		[66]
CRoatan	2017	随机森林；线性回归	结合表达策略预测 sgRNA 的效能预测同源引导修复 Cas9 切割引起的 DSBs 的可能性	序列长度, GC 含量；双链断裂与对应位点的距离	[61]
TKOv3	2017	贝叶斯分析	鉴定必需基因设计全基因组 CRISPR/Cas9 基因文库	评估利用倍数变化分析 sgRNA 靶向；必需基因与非必需基因的分布情况；评估概率分布(贝叶斯因子 BF)	[18]
BAGEL	2016	贝叶斯分析	混合文库筛选鉴定必需基因	评估利用倍数变化分析 sgRNA 靶向；必需基因与非必需基因的分布情况；评估概率分布(贝叶斯因子 BF)	[20]
CRISPRiaDesign	2016	弹性网络线性回归；支持向量回归模型	鉴定 CRISPRa/i 高效率的 sgRNA	序列特征, 位置依赖的序列特征；染色质的状态, 核小体占位率	[8]
CRISPRstrand	2014	基于随机梯度下降的支持向量机	预测 CRISPR 重复序列的方向	ATTGAAAN 重复出现次数；CRISPR 序列核苷酸的组成；序列特定位置的突变；序列折叠成二级结构的倾向性	[19]
H1/H2 library	2018	弹性网络回归算法	筛选 sgRNA 异常值	sgRNA 序列特征	[67]

—：代表该文献未定义该算法工具的名称。

要验证已选特征子集的有效性。

3.5 评估方法

模型和参数确定之后, 通过实验测试评估学习

器的泛化性能。一般地, 通过学习训练数据, 比较模型对输入数据集的预测值与实际值的差异验证模型。常用的模型验证方法是交叉验证(cross validation), 数据的每个子集既是训练集又是验证集。如

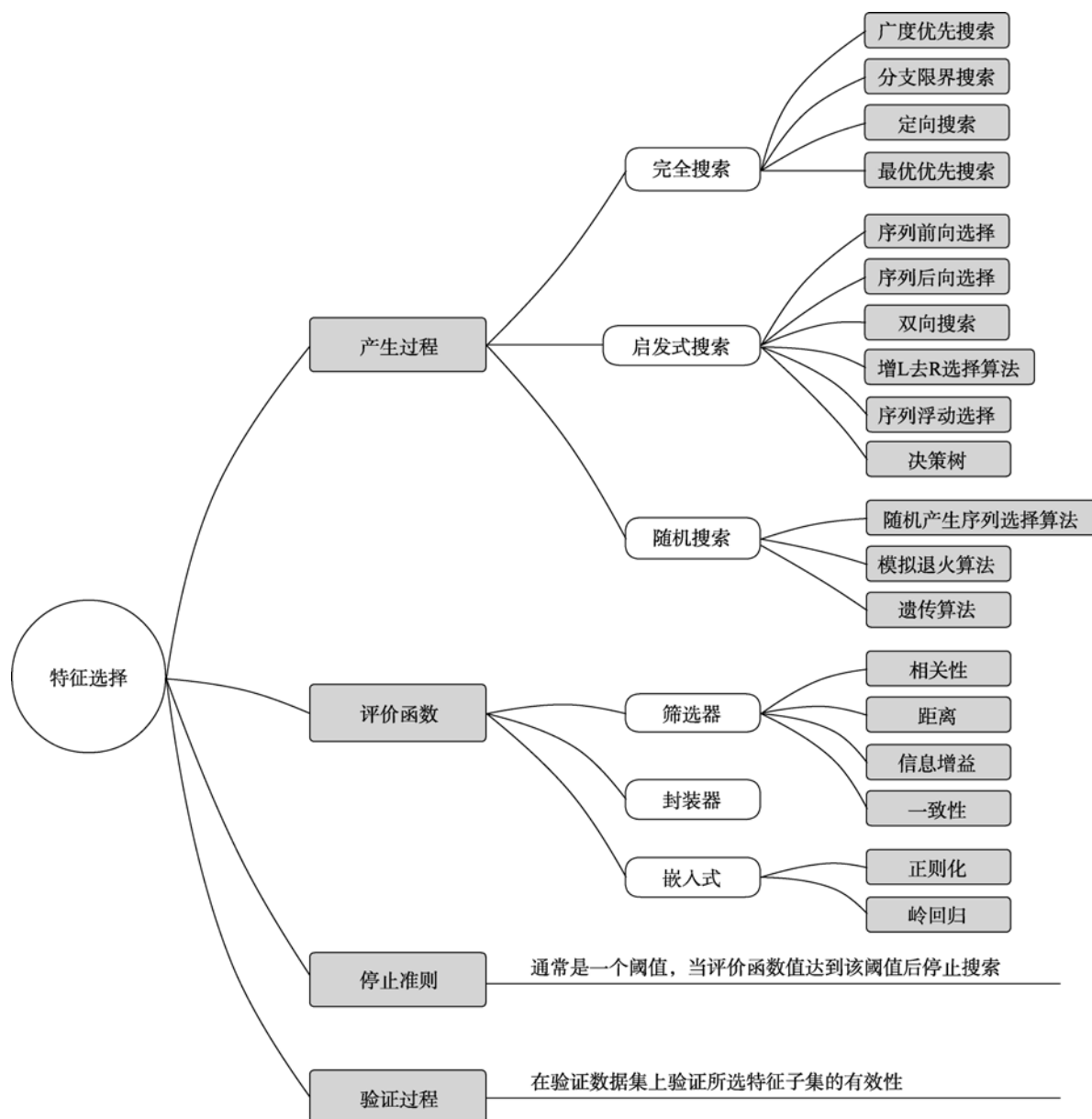


图 3 特征选择方法

Fig. 3 Feature selection methods

三折交叉验证(three-fold cross-validation)^[42]、五折交叉验证(five-fold cross-validation)^[13, 20]、十折交叉验证(ten-fold cross-validation)^[13, 23, 24, 60, 62]、二十折交叉验证(twenty-fold cross-validation)^[13]、嵌套交叉验证(nested cross-validation)^[13, 22]、留一法(leave-one-out, LOO)^[52, 60]等。三折交叉验证将数据分成 3 组, 每一轮依次训练其中的两组数据, 由训练所得参数预测第 3 组数据, 评估模型的准确性。LOO 验证法每次仅测试一个样本, 其他样本用于训练模型。LOO 不受

随机样本划分方式的影响, 评估效果比较准确。然而, 数据集较大的情况下, LOO 训练模型计算成本高。

3.6 性能度量

评估优化 CRISPR/Cas9 系统的机器学习算法学习器的泛化性能, 需要可行的实验评估方法与衡量模型泛化能力的评价体系。通常, 真正例率(true positive rate, TPR)表示正确预测的正样本比例, 假正例率(false positive rate, FPR)表示错误预测的负样本

比例。

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{TN}{TN + FP},$$

其中, TP (true positive) 为真阳性, 表示正确预测的正样本数; TN (true negative) 为真阴性, 表示正确预测的负样本数; FP (false positive) 为假阳性, 表示错误预测的负样本数; FN (false negative) 为假阴性, 表示错误预测的正样本数。准确度 (accuracy, ACC) 和马修相关系数 (Matthew correlation coefficient, MCC) 衡量所有分类器的预测性能:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN},$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

当正样本量与负样本量差别较大时, MCC 能够更公平地反映预测能力。此外, 以“真正例率”为纵轴, “假正例率”为横轴的受试者工作特征曲线 (receiver operating characteristic, ROC) 可作为机器学习算法预测性能的衡量标准^[76], ROC 到达坐标左上角表示所得模型具有较好的性能。

4 常用优化 CRISPR/Cas9 系统的机器学习算法

机器学习在 DNA 编码序列中学习影响 CRISPR/Cas9 系统的模型特征, 为提升该系统基因组编辑效率、最大限度降低脱靶效应提供有力的计算工具。近年来, 机器学习已被应用于预测 sgRNA 的脱靶效应^[13]、优化设计高效 sgRNA^[10, 14, 64]、预测 sgRNA 活性^[17]、设计全基因组 CRISPR/Cas9 基因敲除文库^[18]、鉴定用于 CRISPRi/a 技术的高效率 sgRNA^[8]、鉴定必需基因^[18, 20]等。下面简述几种基于机器学习优化 CRISPR/Cas9 系统的计算方法。

4.1 优化设计高效 sgRNA

设计高效 sgRNA 对提高 CRISPR/Cas9 系统靶点识别特异性有着重要的作用。CRISPR/Cas9 系统可以容忍 sgRNA 与靶 DNA 序列存在一定数量范围的碱基错配, 因此, 靶点可能存在多个候选的 sgRNA。sgRNA 序列特征与其二级结构特征影响靶点切割活

性, 设计 sgRNA 的指导思想是, 由序列的基本特征预测候选 sgRNA 的脱靶效应 (打靶效率), 依据候选 sgRNA 排序结果, 从中选取得分较高的 sgRNA。应用机器学习设计 sgRNA 主要考虑可能优化该系统靶点切割效率的特征, 通过训练模型、调整参数、误差分析, 从而确定影响 sgRNA 打靶效率的特征。Chen 等^[24]基于最小冗余最大相关性 (maximal-relevance minimal-redundancy, mRMR)、增量特征选择算法 (incremental feature selection, IFS)、支持向量机 (support vector machine, SVM) 设计选择影响 CRISPR/Cas9 系统基因组编辑效率与特异性关键因素的算法, 有效地剔除冗余特征。

CRISTA 方法^[52]基于随机森林 (random forest) 与回归模型 (regression model), 考虑了 DNA 凸起 (DNA bulge)/RNA 凸起 (RNA bulge) 对 sgRNA 编辑效率的影响。DNA 凸起 (RNA 凸起) 指脱靶位点 DNA 序列长度比 sgRNA 多 (少) 若干个碱基, 通过形成凸起完成碱基的正确配对。CRISTA 方法考虑 DNA 凸起, 结合基因组核苷酸含量、sgRNA 热力学特征、sgRNA 与靶 DNA 序列元素相似性等, 评估 sgRNA 切割基因组位点的倾向性。该作者改进 Needleman-Wunch 双序列比对算法。在 sgRNA 与目标 DNA 序列比对中, 最多容忍 3 个单碱基的间隙 (gap), 这有利于高效预测潜在的脱靶位点。由于最多允许 3 个单碱基的间隙, 故将长度为 20 bp 的靶 DNA 序列缩短 (延伸) 3 个碱基。随机选择其中 7 个长度为 17~23 bp 的靶 DNA 序列与相应 sgRNA 序列做序列比对, 选取最高分数的 DNA 序列作为靶序列。最后, 计算 sgRNA 对靶 DNA 序列切割倾向性与序列比对分数的最大均方皮尔逊相关系数 (maximal averaged squared Pearson correlation coefficient), 确定最优模型参数。研究表明, DNA 凸起/RNA 凸起是 CRISPR/Cas9 系统的组成部分 (integral part)^[52]。在该作者测试的数据集中, 凸起约占 20%。大多数凸起 sgRNA 对目标 DNA 序列的切割效率较低, 少数可达到中等切割效率。

4.2 预测 sgRNA 的编辑特异性

CRISPRpred 方法^[60]基于支持向量机预测 sgRNA

对目标DNA序列的打靶活性。利用随机森林算法计算平均下降基尼系数(Gini index),选取所有影响sgRNA编辑特异性的特征。基尼系数表示模型的不纯度。基尼系数越小,不纯度越低,特征越好。决策树选择基尼系数增益值最大的特征,作为节点的分裂条件^[77]。CRISPRpred方法利用十折交叉验证,计算均方根误差(root-mean-square error, RMSE)验证特征之间的相关性。通过设置参数(学习率、随机森林树的数量)验证模型防止过拟合。利用已选特征的不同组合分析FC-RES数据集。该作者利用留一法交叉验证计算评估矩阵,由ROC曲线、PR (precision-recall)曲线评估模型的性能。研究表明,考虑氨基酸的切割位置、肽百分比有利于估计sgRNA的编辑特异性。

Kuan等^[23]根据寡核苷酸设计(oligonucleotide design)^[78]与核小体占用模型^[79]的先验知识,基于机器学习评估影响sgRNA对靶DNA序列切割效率的特征,提出设计高编辑效率的sgRNA的计算方法。该作者考虑可能影响sgRNA靶向目标DNA序列切割效率的特征:(1)位置依赖的单核苷酸(position-dependent mono-nucleotide, PD Mono),如sgRNA第一个符号为A记为“A_1”; (2)位置独立的单核苷酸(position-dependent dinucleotide, PD Dinuc)特征,如sgRNA序列A的数目;(3)单核苷酸与二核苷酸的频率;(4)sgRNA与目标DNA序列的比对得分;(5)热力学特征及二级结构、理化性质(physiochemical properties); (6)由PseKNC模型生成伪k-元组核苷酸特征。结合逻辑回归、随机森林分析每个特征对sgRNA编辑效率的贡献。研究发现,T和TT二核苷酸的频率与sgRNA的编辑效率呈强负相关。TT二核苷酸在移位方面灵活性较小,在丰度较高的区域sgRNA的编辑效率较低。结合位置依赖的二核苷酸特征优化ROC曲线下的面积(area under ROC curve, AUC)的性能预测sgRNA的活性。染色质可达性(chromatin accessibility)已被证明是dCas9-sgRNA基因组结合的主要决定因素^[43]。与染色质重塑、染色质可达性相关的表观遗传标记包括DNase I超敏感位点、转录因子结合、DNA甲基化和组蛋白修饰。综合考虑核苷酸组成、染色质结构特征与sgRNA表达载体依赖性水平的特征有利于预测sgRNA的特异性。

4.3 评估sgRNA的脱靶效应

Elevation方法^[13]基于CFD^[22]分三步评估给定sgRNA的脱靶效应。首先,利用已知数据训练第一层机器学习模型。根据目标DNA序列与sgRNA至多存在N个碱基不匹配的原则,筛选sgRNA在基因组所有可能的靶点。接着,为每个潜在的靶点评分,量化sgRNA与靶序列的脱靶活性。最后,评估其脱靶活性,计算每个sgRNA的综合得分。该方法运用二层回归模型,第一层模型(单个碱基不匹配)考虑位置依赖的sgRNA与目标DNA序列碱基不匹配的位置、不匹配核苷酸类型、转换突变、颠换突变等,计算基尼系数分析特征的重要性;第二层模型(多错配组合)计算sgRNA与目标DNA序列碱基不匹配得分与第一层单个碱基不匹配得分之和。Listgarten等^[13]利用美国哈佛医学院和马萨诸塞州总医院的公开数据集对第二层模型进行训练,将第一层模型加以细化,推广至碱基错配数目大于1的靶标区域。基于机器学习方法计算基因组可能发生脱靶的各个区域的概率计算脱靶分值。Elevation算法计算sgRNA两类脱靶分值:该sgRNA在特定靶标区域内的脱靶分值及其在所有可能的靶区域内对应的脱靶分值。

4.4 CRISPR遗传筛选

CRISPR/Cas9为功能性基因筛选(functional genetic screening)提供高效简便的技术支持。利用CRISPR可以通过诱导基因突变进行功能缺失型(loss of function)筛选,也可以激活转录进行功能获得型(gain of function)筛选。CRISPR基因筛选可以作用于基因编码区(coding region),也可以靶向基因非编码区(non-coding region)^[4]。基于CRISPR/Cas9功能基因筛选技术在细胞生存必需基因鉴定^[18, 20]、潜在治疗靶标筛选^[6]、肿瘤转移^[80]等方面已取得重要的研究进展。

近年来,CRISPR/Cas9在基因扰动(genetic perturbation)领域已被应用于基因编辑、基因的上调和下调表达,主要包括基因敲除(gene knockout, KO)、基因敲入(gene knockin, KI)、CRISPR干扰/CRISPR激活^[12, 81, 82]。CRISPRi将Cas9突变失活(dead Cas9, dCas9),使其无法切割DNA双链,再与转录抑制因

子(transcription repression factor)作用,从而实现在 sgRNA 的指导下抑制特定基因的表达^[83]。CRISPRa 利用 dCas9 与转录激活因子(transcriptional activator)作用,上调靶基因的表达水平,用于功能获得型筛选^[84]。运用 CRISPR/Cas9 系统进行基因扰动,sgRNA 的特异性对靶基因筛选至关重要。合理构建包含靶点信息的 sgRNA 文库,可以有效地提高靶基因筛选的效率,实现最大限度地降低脱靶效应^[85]。CRISPR 筛选(CRISPR screen)常用混合文库(pooled-library),将合成混合寡核苷酸库装入逆转录病毒(如慢病毒)载体感染宿主细胞,再将文库序列整合至基因组中表达,检测细胞生长表型^[36]。混合文库通常采用定向筛选,包括正向筛选(positive screening)与逆向筛选(negative screening)。正向筛选通过施加一个强的选择压力,经过文库扰动,野生型基因致死,获得抗性的克隆细胞存活,从而鉴定可产生抗性的基因^[86]。逆向筛选用于鉴定细胞必需基因^[87]。

机器学习在 CRISPR 遗传筛选领域有着巨大的潜能。机器学习已逐渐应用于设计高编辑效率 sgRNA 文库^[67]、鉴定必需基因^[18, 20]、结合基因表达策略预测 sgRNA 的编辑效率^[61]、鉴定 CRISPRa/i 高效率的 sgRNA^[8, 19]等。

4.4.1 设计全基因组 sgRNA 文库

基因筛选 sgRNA 文库构建以全基因组基因为靶点,针对每个靶点设计效率较高的一组 sgRNA。设计全基因组 sgRNA 文库仍存在以下问题:(1) sgRNA 异常值、sgRNA 与靶向相同基因的其他 sgRNA 活性存在较大的差异;(2) CRISPR/Cas9 系统中,间隔区(spacer)长度可能不同^[77, 88],仅对单个向导(靶序列)分析其最佳间隔区长度;(3)间隔区长度与信噪比(signal-to-noise ratio, SNR)的关系尚不明确^[67]。基于机器学习算法设计全基因组 sgRNA 文库的指导思想是,考虑影响 sgRNA 编辑效率的特征,利用机器学习算法进行训练并评分^[85, 89],根据得分选取高效的 sgRNA 构建大规模文库。

Chen 等^[67]基于 MAGeCK-VISPR 模型^[90]识别 sgRNA 异常值(sgRNA outlier),利用弹性网络回归算法(elastic-net regression)^[42]提取 sgRNA 序列特征,根据 sgRNA 是否为异常值分别赋值为 1 或 0。由于

不同 sgRNA 的切割活性与修复效率、局部染色质结构与潜在的脱靶效应不同。在基因筛选中,不同 sgRNA 靶向同一个基因,将产生不同的细胞生长表型与选择水平^[39, 91, 92]。该作者分析已有文库发现,在非种子区域 G 较高的 sgRNA 具有较强的脱靶活性,导致较强的异常表型。将 sgRNA 靶向多个非必需基因作为阴性对照,能够降低假阳性。比较不同长度(18, 19, 20 个碱基) sgRNA 的切割效率及信噪比,得到长度为 19 个碱基的 sgRNA 具有最高的切割效率,提示其在基因敲除效应中更稳定(潜在的脱靶效应小)。基于此,该作者提出一种全新的全基因组 CRISPR/Cas9 筛选文库(H1/H2)与 Brunello 文库^[22]、TKO 文库^[93]、Ong 文库^[94]比较,该文库 sgRNA 异常值比例最小。比较 GeCKOv2 与 TKO 文库 H1/H2、Brunello 与 Ong 文库对已知必需基因的识别性能更优。

4.4.2 鉴定必需基因

BAGEL (Bayesian analysis of gene Essentiality)方法^[20]基于贝叶斯分析基因敲除筛选,识别混合文库筛选中的必需基因。BAGEL 首先估计所有靶向必需基因、靶向非必需基因训练集的 sgRNA 的表达倍数变化的分布(distribution of fold changes)。其次,使用核密度估计(kernel density estimation)sgRNA 靶向参考必需基因与非必需基因的似然函数,计算贝叶斯因子(Bayes factor, BF)。Hart 等^[20]将必需基因、非必需基因数据集作为测试集,计算 PR 曲线评估筛选性能。研究表明,利用 BAGEL 分析基因敲除筛选,能够灵敏、准确地鉴定适应性基因(fitness gene),且大大降低计算时间。运用 BAGEL 鉴定人类细胞系混合库基因敲除筛选 2000 个适应性基因,错误发现率(false discovery rate, FDR)为 5%。而且,BAGEL 对不同平台文库筛选具有高灵敏性与特异性。

CRISPR/Cas9 为哺乳动物细胞高通量功能基因筛选提供技术支持。人类蛋白编码基因与病毒载体表达的 CRISPR sgRNA 混合文库已应用于人类多种癌症细胞与永生细胞系基因敲除^[18]。研究表明,CRISPR 筛选比混合 shRNA 筛选更敏感^[95],然而,CRISPR 文库设计和实验方案中存在显著的偏差。Hart 等^[18]报道了利用 CRISPR/Cas9 系统评估与设计全基因组遗传筛选。该作者使用来自 3 个研究组不

同基因组规模的 sgRNA 文库分析人类细胞系 17 个基因敲除筛选,使用 BAGEL 算法^[20]鉴定必需基因,将已定义的人类核心必需基因由 360 个扩展至 684。根据扩展的核心必需基因参考基(CEG2)以及来自 6 个 CRISPR 敲除筛选的经验数据,设计优化序列的 sgRNA 文库(Toronto KnockOut version 3.0, TKOv3)^[18]。结合 CEG2 参考基的优化 TKOv3 文库,为评估人类细胞系基因敲除筛选提供一个高效的平台^[18]。

4.4.3 鉴定 CRISPRi/a 高效率的 sgRNA

CRISPRi/a 主要靶向基因启动子区,设计 sgRNA 所考虑的特征与 CRISPR 基因敲除(CRISPR KO)系统稍有不同。与 CRISPR/Cas9 基因敲除类似,CRISPRi/a 系统序列间隔区也具有嘌呤偏好性。与 CRISPR/Cas9 系统相比,CRISPRi/a 具有独特的效应区(effector domain),而且,该结构域在基因扰动中发挥关键作用^[42, 96]。由于缺乏足够的数据来源,仅有少数工具用于 CRISPRi/a 系统设计 sgRNA (如 CRISPR-ERA)^[97]。序列特征分析将有助于提高对 sgRNA 编辑效率的预测性能^[85]。

Horlbeck 等^[98]报道了核小体能够直接阻断 CRISPR/Cas9 接近目标 DNA。CRISPRi/a 需要持续的 dCas9 与 DNA 相结合,考虑核小体占位有利于预测 CRISPRi/a 高编辑效率的 sgRNA^[8]。Horlbeck 等^[8]基于机器学习结合染色质、sgRNA 序列等特征,设计 CRISPRi/a Design 工具用于预测 CRISPRi/a 高效率的 sgRNA。CRISPRi 的活性与靶点与转录起始点的距离之间具有周期性与不对称性^[98]。基于此,该作者首先运用支持向量回归(support vector regression, SVR)拟合 sgRNA 位置特征,从而预测靶位点特征的连续函数。接着,利用弹性网络线性回归对 30 个 CRISPRi 筛选的数据进行训练分析,结合靶位点距离转录起始点、核小体占位、位置依赖的二核苷酸、位置依赖的单核苷酸、sgRNA 二级结构、靶位点染色质的易接近性、sgRNA 的长度等特征,预测 sgRNA 的编辑活性分数。最后,利用该算法设计人类(hCRISPRi-v2)和小鼠基因组 CRISPRi/a 文库(version 2)。K562 细胞必需基因 CRISPRi 筛选实验表明,大多数 sgRNA 具有较高的活性。PR 曲线分析显示,采用一个紧凑的 sgRNA 基因文库检测超过 90%的

必需基因假阳性最小。CRISPRi/a 作为功能缺失型筛选与功能获得型筛选研究的主要工具,能够为识别 Cas9 靶点提供一个通用的工具。

4.4.4 预测 sgRNA 的活性

Croatan 方法^[61]结合随机森林与 sgRNA 的表达策略优化 CRISPR/Cas9 系统基因敲除效率。研究表明,利用多个 RNA 聚合酶可以促进 sgRNA 的独立性,优化基因的表达策略^[99]。Cpf1 可以靶向独立表达 sgRNA 的 crRNA 阵列的细胞多个靶点^[100]。上述策略有助于研究细胞中单基因敲除的位点,但是,Cas9 同时靶向目标序列多个位点可能导致更大的功能性后果(functional consequence)。结合选择算法与基因表达策略将有利于选择高效能的 sgRNA。Erard 等^[61]基于随机森林设计 Croatan 算法用于预测 sgRNA 切割效率,结合基因表达策略,可用于人类细胞系单个基因敲除、多个基因敲除。Croatan 结合核苷酸组合、移码突变(frameshift mutation, FSM)可能性与靶点是否在编码区特征,训练 Doench 等^[43]与 Chari 等^[66]的数据集选择高编辑效率的 sgRNA 识别目标靶点。研究表明,靶序列的保守性(target conservation)与靶序列侧翼同源序列(target-flanking homologous sequence)影响 sgRNA 的切割效率。而且,Cas9 作用于多个靶位点将增强功能影响。当两个独立的 sgRNA 同时引导 Cas9 靶向目标基因,编辑效率将显著提高。对于每个靶位点,依据预测分数较高的两个 sgRNA 设计文库。基于此,该作者设计与建立全基因组 CRISPR 阵列文库(arrayed CRISPR library),该文库可用多重或阵列格式进行单个(合并)遗传筛选。

5 存在的问题与潜在的解决方案

CRISPR/Cas9 系统基因组编辑效率和特异性受诸多因素的影响。自 2014 年起,已有多种基于机器学习的计算方法用于优化该系统。然而,这些方法对该系统基因组编辑效率和脱靶效应的研究尚不够深入,研究结果并不一致。这反映在不同计算方法数据整合、特征选择、sgRNA 的编辑活性的评价标准不一致。机器学习优化 CRISPR/Cas9 系统依赖先

验知识,需要大量学习已知实验数据。不同 sgRNA 设计和脱靶效应评估软件的操作平台、参数设置与评估指标不同,输出数据格式也不一致。如 Digenome-Seq^[101, 102]未提供活体实验的基因编辑切割频率的数据。此外,由于算法原理不同,运用不同的计算的方法得到的研究结论不完全一致。如针对影响 sgRNA 切割效率的特征这一问题,Haussler 等^[103]发现凸起极少发生,其切割效率忽略不计。而 Abadi 等^[52]测试的数据集中,凸起约占 20%。而且大多数凸起 sgRNA 对靶 DNA 序列的切割效率较低,少数凸起 sgRNA 达到中等切割效率水平。

运用机器学习预测 CRISPR/Cas9 系统基因组编辑效率主要分析序列的特征。有些计算方法为了降低计算成本,忽略一些重要的特征。如 Elevation 方法^[13]第二层整合分数模型,考虑 sgRNA 与目标 DNA 序列碱基错配类型的组合容易产生组合爆炸。为了简化模型,该方法忽略插入/缺失,主要考虑 sgRNA 与目标 DNA 序列碱基不匹配特征。另外,与表观遗传修饰有关的特征可能影响 CRISPR/Cas9 系统的性能。然而,大多数计算方法并未考虑这方面的特征。

单碱基编辑技术(base editor)^[104, 105]是基于 CRISPR 系统的新型靶基因定点修饰技术,它不需要产生 DNA 双链断裂及 DNA 模板,能够对基因组特定碱基进行高效的替换。单碱基编辑 PAM 序列主要为 NGG,能够进行修饰的序列较少^[106]。目前,单碱基编辑技术的作用机制有待深入探究。David^[106]等发现 xCas9 酶在单碱基编辑领域得到更广泛的 PAM,而且脱靶效应远低于 spCas9,但其作用机制尚不明确。运用机器学习学习先验知识,可以帮助科研工作者寻找影响单碱基编辑效率的关键特征,从而提高单碱基编辑效率,降低脱靶效应。运用机器学习优化单碱基编辑技术将为实现基因的功能缺失突变、基因的表达调控方面、植物基因组功能解析和作物遗传改良及新品种培育提供了重要技术支撑^[104, 106, 107]。

机器学习应用于优化 CRISPR/Cas9 系统尚存在一些问题有待进一步解决。研究者在寻找弥补不足的方法的同时,也在不断拓展该技术的应用领域。深度学习(deep learning)可以避免耗时费力的特征工程提取过程,运用深度学习分析 CRISPR/Cas9 系统是一个创新可行的思路。DeepCpf1 算法^[14]基于卷积

神经网络(convolutional neural network, CNN)结合染色质易接近性特征,预测 CRISPR/Cpf1 系统 sgRNA 的编辑活性。该方法比传统机器学习方法(如 L1/L2 正则线性回归、梯度上升回归树)预测性能更优。随着研究的不断深入,这些问题将会得到解决,进而推动功能基因组学、基因表观遗传调控、疾病治疗等领域的发展。

6 结语与展望

CRISPR/Cas9 系统是由细菌和古细菌等微生物特有的获得性免疫系统发展起来的基因组编辑技术,能够对基因组特定位点进行基因敲除、基因敲入、DNA 大片段删除、转录调控等遗传操作。CRISPR/Cas9 技术凭借成本低廉、效率高和易操作性等优点在基因工程领域具有很好的发展潜力。CRISPR/Cas9 技术处于初步的研究阶段,影响该系统基因组编辑效率和脱靶效应的机制基本明确。但是,仍存在很多问题有待进一步优化,如提高基因组编辑效率与编辑特异性等。如何提高同源重组修复的效率,如何降低非同源重组修复造成的非预期的突变,是当前利用 CRISPR/Cas9 技术进行基因治疗所面临的技术难题。

机器学习为优化 CRISPR/Cas9 系统所面临的问题提供创新的解决思路。通过对不同实验数据进行整合分析,基于机器学习建立数学模型分析影响 CRISPR/Cas9 系统编辑效率、编辑特异性的特征,有助于研究者们更深入地理解该系统的作用机制。机器学习在 CRISPR/Cas9 系统脱靶效应评估、预测 sgRNA 的活性、设计高效 sgRNA、基因敲除、高通量功能基因筛选等研究领域有着日渐广泛的应用。尽管机器学习应用于优化 CRISPR/Cas9 技术尚处于研究发展阶段,但已显现出广阔的应用前景。通过不断改进,机器学习、深度学习应用于优化 CRISPR/Cas9 系统的应用将成为研究热点。今后,机器学习/深度学习优化 CRISPR/Cas9 系统技术将会为临床医师和科研人员提供更多有价值的信息,进而在基础科学研究、分子生物学研究和基因治疗等诸多领域产生深远的影响。

参考文献(References):

- [1] Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini L, Zhang F. Multiplex genome engineering using CRISPR/Cas systems. *Science*, 2013, 339(6121): 819–823. [DOI]
- [2] Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat Biotechnol*, 2013, 31(3): 233–239. [DOI]
- [3] Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, 2014, 157(6): 1262–1278. [DOI]
- [4] Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK, Sander JD. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol*, 2013, 31(9): 822–826. [DOI]
- [5] Pattanayak V, Lin S, Guilinger JP, Ma E, Doudna JA, Liu DR. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat Biotechnol*, 2013, 31(9): 839–843. [DOI]
- [6] Wong N, Liu W, Wang X. WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. *Genome Biol*, 2015, 16: 218. [DOI]
- [7] Hinz JM, Laughery MF, Wyrick JJ. Nucleosomes inhibit Cas9 endonuclease activity *in vitro*. *Biochemistry*, 2015, 54(48): 7063–7066. [DOI]
- [8] Horlbeck MA, Gilbert LA, Villalta JE, Adamson B, Pak RA, Chen Y, Fields AP, Park CY, Corn JE, Kampmann M, Weissman JS. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife*, 2016, 5: e19760. [DOI]
- [9] Lee CM, Davis TH, Bao G. Examination of CRISPR/Cas9 design tools and the effect of target site accessibility on Cas9 activity. *Exp Physiol*, 2017, 103(4): 456–460. [DOI]
- [10] Isaac RS, Jiang FG, Doudna JA, Lim WA, Narlikar GJ, Almeida R. Nucleosome breathing and remodeling constrain CRISPR-Cas9 function. *Elife*, 2016, 5: e13450. [DOI]
- [11] Kosicki M, Tomberg K, Bradley A. Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. *Nat Biotechnol*, 2018, 36(8): 765–771. [DOI]
- [12] Goldberg DE. Genetic Algorithms in Search, Optimization and Machine Learning. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989. [DOI]
- [13] Listgarten J, Weinstein M, Kleinstiver BP, Sousa AA, Joung JK, Crawford J, Gao K, Hoang L, Elibol M, Doench JG, Fusi N. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nat Biomed Eng*, 2018, 2(1): 38–47. [DOI]
- [14] Kim HK, Min S, Song M, Jung S, Choi JW, Kim Y, Lee S, Yoon S, Kim HH. Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. *Nat Biotechnol*, 2018, 36(3): 239–241. [DOI]
- [15] Lin Y, Cradick TJ, Brown MT, Deshmukh H, Ranjan P, Sarode N, Wile BM, Vertino PM, Stewart FJ, Bao G. CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res*, 2014, 42(11): 7473–7485. [DOI]
- [16] Najm FJ, Strand C, Donovan KF, Hegde M, Sanson KR, Vaimberg EW, Sullender ME, Hartenian E, Kalani Z, Fusi N, Listgarten J, Younger ST, Bernstein BE, Root DE, Doench JG. Orthologous CRISPR–Cas9 enzymes for combinatorial genetic screens. *Nat Biotechnol*, 2017, 36(2): 179–189. [DOI]
- [17] Kescu C, Arslan S, Singh R, Thorpe J, Adli M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat Biotechnol*, 2014, 32(7): 677–683. [DOI]
- [18] Hart T, Tong A, Chan K, van Leeuwen J, Seetharaman A, Aregger M, Chandrashekhara M, Hustedt N, Seth S, Noonan A, Habsid A, Sizova O, Nedyalkova L, Climie R, Lawson K, Sartori MA, Alibai S, Tieu D, Masud S, Mero P, Weiss A, Brown KR, Ušaj M, Billmann M, Rahman M, Costanzo M, Myers CL, Andrews B, Boone C, Durocher D, Moffat J. Evaluation and design of genome-wide CRISPR/Cas9 knockout screens. *bioRxiv*. 2017, 7(8): 2719–2727. [DOI]
- [19] Alkhnbashi OS, Costa F, Shah SA, Garrett RA, Saunders SJ, Backofen R. CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics*, 2014, 30(17): i489–i496. [DOI]
- [20] Hart T, Moffat J. BAGEL: a computational framework for identifying essential genes from pooled library screens. *BMC Bioinformatics*, 2016, 17: 164. [DOI]
- [21] Kim HK, Song M, Lee J, Menon AV, Jung S, Kang YM, Choi JW, Woo E, Koh HC, Nam JW, Kim H. In vivo high-throughput profiling of CRISPR-Cpf1 activity. *Nat Methods*, 2017, 14(2): 153–159. [DOI]
- [22] Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg

- EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, Virgin HW, Listgarten J, Root DE. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, 2016, 34(2): 184–191. [DOI]
- [23] Kuan PF, Powers S, He S, Li K, Zhao X, Huang B. A systematic evaluation of nucleotide properties for CRISPR sgRNA design. *BMC Bioinformatics*, 2017, 18(1): 297. [DOI]
- [24] Chen L, Wang SP, Zhang YH, Li JR, Xing ZH, Yang J, Huang T, Cai YD. Identify key sequence features to improve CRISPR sgRNA efficacy. *IEEE Access*, 2017, 5: 26582–26590. [DOI]
- [25] Shah SA, Vestergaard G, Garrett RA. CRISPR/Cas and CRISPR/Cmr Immune Systems of Archaea. Springer Vienna, 2012: 163–181. [DOI]
- [26] Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J Bacteriol*, 1987, 169(12): 5429–5433. [DOI]
- [27] Jansen R, Embden JD, Gaastra W, Schouls LM. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol*, 2002, 43(6): 1565–1575. [DOI]
- [28] Garneau JE, Dupuis M-È, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadán AH, Moineau S. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, 2010, 468(7320): 67–71. [DOI]
- [29] Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. *Science*, 2010, 327(5962): 167–170. [DOI]
- [30] Mojica FJM, DíezVillaseñor C, García-Martínez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, 2009, 155(pt 3): 733–740. [DOI]
- [31] Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase . *Nature*, 2011, 471(7340): 602–607. [DOI]
- [32] Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N, Ishitani R, Zhang F, Nureki O. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, 2014, 156(5): 935–949. [DOI]
- [33] Lu XJ, Xue HY, Ke ZP, Chen JL, Ji LJ. CRISPR-Cas9: a new and promising player in gene therapy. *J Med Genet*, 2015, 52(5): 289–296. [DOI]
- [34] Rouet P, Smih F, Jasin M. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol Cell Biol*, 1994, 14(12): 8096–8106. [DOI]
- [35] Rouet P, Smih F, Jasin M. Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. *Proc Natl Acad Sci USA*, 1994, 91(13): 6064–6068. [DOI]
- [36] Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 2012, 337(6096): 816–821. [DOI]
- [37] Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*, 2014, 507(7490): 62–67. [DOI]
- [38] Zhang Y, Ge X, Yang F, Zhang L, Zheng J, Tan X, Jin ZB, Qu J, Gu F. Comparison of non-canonical PAMs for CRISPR/Cas9-mediated DNA cleavage in human cells. *Sci Rep*, 2014, 4: 5405. [DOI]
- [39] Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O, Cradick TJ, Marraffini LA, Bao G, Zhang F. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol*, 2013, 31(9): 827–832. [DOI]
- [40] Kleinstiver BP, Prew MS, Tsai SQ, Topkar VV, Nguyen NT, Zheng Z, Gonzales AP, Li Z, Peterson RT, Yeh JR, Aryee MJ, Joung JK. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature*, 2015, 523(7561): 481–485. [DOI]
- [41] Mali P, Aach J, Stranges PB, Esvelt KM, Moosburner M, Kosuri S, Yang L, Church GM. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol*, 2013, 31(9): 833–838. [DOI]
- [42] O'Geen H, Henry IM, Bhakta MS, Meckler JF, Segal DJ. A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. *Nucleic Acids Res*, 2015, 43(6): 3389–3404. [DOI]
- [43] Wu X, Scott DA, Kriz AJ, Chiu AC, Hsu PD, Dadon DB, Cheng AW, Trevino AE, Konermann S, Chen S, Jaenisch R, Zhang F, Sharp PA. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat Biotechnol*, 2014, 32(7): 670–676. [DOI]
- [44] Bae S, Kweon J, Kim HS, Kim JS. Microhomology-

- based choice of Cas9 nuclease target sites. *Nat Methods*, 2014, 11(7): 705–706. [DOI]
- [45] Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, Sullender M, Ebert BL, Xavier RJ, Root DE. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol*, 2014, 32(12): 1262–1267. [DOI]
- [46] Moreno-Mateos MA, Vejnar CE, Beaudoin JD, Fernandez JP, Mis EK, Khokha MK, Giraldez AJ. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting *in vivo*. *Nat Methods*, 2015, 12(10): 982–988. [DOI]
- [47] Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, 2014, 343(6166): 80–84. [DOI]
- [48] Cho SW, Kim S, Kim Y, Kweon J, Kim HS, Bae S, Kim JS. Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res*, 2014, 24(1): 132–141. [DOI]
- [49] Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat Biotechnol*, 2014, 32(3): 279–284. [DOI]
- [50] Koch B, Nijmeijer B, Kueblbeck M, Cai Y, Walther N, Ellenberg J. Generation and validation of homozygous fluorescent knock-in cells using CRISPR-Cas9 genome editing. *Nat Protoc*, 2018, 13(6): 1465–1487. [DOI]
- [51] Ran FA, Hsu Patrick D, Lin CY, Gootenberg Jonathan S, Konermann S, Trevino AE, Scott David A, Inoue A, Matoba S, Zhang Y. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, 2013, 154(6): 1380–1389. [DOI]
- [52] Abadi S, Yan WX, Amar D, Mayrose I. A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. *PLoS Comput Biol*, 2017, 13(10): e1005807. [DOI]
- [53] Xie S, Shen B, Zhang C, Huang X, Zhang Y. sgRNAs9: a software package for designing CRISPR sgRNA and evaluating potential off-target cleavage sites. *PLoS One*, 2014, 9(6): e100448. [DOI]
- [54] MacPherson CR, Scherf A. Flexible guide-RNA design for CRISPR applications using Protospacer Workbench. *Nat Biotechnol*, 2015, 33(8): 805–806. [DOI]
- [55] Ma M, Ye AY, Zheng W, Kong L. A guide RNA sequence design platform for the CRISPR/Cas9 system for model organism genomes. *Biomed Res Int*, 2013, 2013: 270805. [DOI]
- [56] Guilinger JP, Thompson DB, Liu DR. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nat Biotechnol*, 2014, 32(6): 577–582. [DOI]
- [57] Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc*, 2013, 8(11): 2281–2308. [DOI]
- [58] Kleinstiver BP, Pattanayak V, Prew MS, Tsai SQ, Nguyen NT, Zheng Z, Joung JK. High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature*, 2016, 529(7587): 490–495. [DOI]
- [59] Frock RL, Hu J, Meyers RM, Ho Y-J, Kii E, Alt FW. Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat Biotechnol*, 2015, 33(2): 179–186. [DOI]
- [60] Rahman MK, Rahman MS. CRISPRpred: A flexible and efficient tool for sgRNAs on-target activity prediction in CRISPR/Cas9 systems. *PLoS One*, 2017, 12(8): e0181943. [DOI]
- [61] Erard N, Knott SRV, Hannon GJ. A CRISPR Resource for Individual, Combinatorial, or Multiplexed Gene Knockout. *Mol Cell*, 2017, 67(6): 1080. [DOI]
- [62] Chari R, Yeo NC, Chavez A, Church GM. sgRNA Scorer 2.0: A species-independent model to predict CRISPR/Cas9 activity. *ACS Synth Biol*, 2017, 6(5): 902–904. [DOI]
- [63] Ma J, Koster J, Qin Q, Hu S, Li W, Chen C, Cao Q, Wang J, Mei S, Liu Q, Xu H, Liu XS. CRISPR-DO for genome-wide CRISPR design and optimization. *Bioinformatics*, 2016, 32(21): 3336–3338. [DOI]
- [64] Prykhodzhiy SV, Rajan V, Gaston D, Berman JN. CRISPR multitargeter: a web tool to find common and unique CRISPR single guide RNA targets in a set of similar sequences. *PLoS One*, 2015, 10(3): e0119372. [DOI]
- [65] Xu H, Xiao T, Chen C-H, Li W, Meyer CA, Wu Q, Wu D, Cong L, Zhang F, Liu JS, Brown M, Liu XS. Sequence determinants of improved CRISPR sgRNA design. *Genome Res*, 2015, 25(8): 1147–1157. [DOI]
- [66] Chari R, Mali P, Moosburner M, Church GM. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat Methods*, 2015, 12(9): 823–826. [DOI]
- [67] Chen CH, Xiao T, Xu H, Jiang P, Meyer CA, Li W, Brown M, Liu XS. Improved design and analysis of CRISPR knockout screens. *Bioinformatics*, 2018, doi: 10.1093/bioinformatics/bty450. [DOI]

- [68] Box GEP, Cox DR. An Analysis of Transformations. *J Roy Statist Soc Ser B*, 1964, 26(2): 211–252. [DOI]
- [69] Ran FA, Cong L, Yan WX, Scott DA, Gootenberg JS, Kriz AJ, Zetsche B, Shalem O, Wu XB, Makarova KS, Makarova KS, Koonin E, Sharp PA, Zhang F. *In vivo* genome editing using *Staphylococcus aureus* Cas9. *Nature*, 2015, 520(7546): 186–191. [DOI]
- [70] Slaymaker IM, Gao L, Zetsche B, Scott DA, Yan WX, Zhang F. Rationally engineered Cas9 nucleases with improved specificity. *Science*, 2016, 351(6268): 84–88. [DOI]
- [71] Tsai SQ, Zheng Z, Nguyen NT, Liebers M, Topkar VV, Thapar V, Wyvekens N, Khayter C, Iafrate AJ, Le LP, Aryee MJ, Joung JK. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol*, 2015, 33(2): 187–197. [DOI]
- [72] Hilton IB, D'ippolito AM, Vockley CM, Thakore PI, Crawford GE, Reddy TE, Gersbach CA. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat Biotechnol*, 2015, 33(5): 510–517. [DOI]
- [73] Friedman JH. On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data Min Knowl Disc*, 1997, 1(1): 55–77. [DOI]
- [74] Grabczewski K, Jankowski N. Mining for complex models comprising feature selection and classification. *Feat Extrac*, 2004, 207: 473–489. [DOI]
- [75] Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res*, 2003, 3: 1157–1182. [DOI]
- [76] Faraggi D, Reiser B. Estimation of the area under the ROC curve. *Stat Med*, 2002, 21(20): 3093–3106. [DOI]
- [77] Robnik-Šikonja M. Improving Random Forests. *Lect Not Comput Sci*, 2004, 3201: 359–370. [DOI]
- [78] Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. A model-based background adjustment for oligonucleotide expression arrays. *Publ Amer Stat Assoc*, 2004, 99(468): 909–917. [DOI]
- [79] Chen W, Lin H, Feng PM, Ding C, Zuo YC, Chou KC. iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. *PLoS One*, 2012, 7(10): e47843. [DOI]
- [80] Chen S, Sanjana NE, Zheng K, Shalem O, Lee K, Shi X, Scott DA, Song J, Pan JQ, Weissleder R, Lee H, Zhang F, Sharp PA. Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell*, 2015, 160(6): 1246–1260. [DOI]
- [81] Sander JD, Joung JK. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol*, 2014, 32(4): 347–355. [DOI]
- [82] Graham DB, Root DE. Resources for the design of CRISPR gene editing experiments. *Genome Biol*, 2015, 16: 260. [DOI]
- [83] Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, Stern-Ginossar N, Brandman O, Whitehead EH, Doudna JA, Lim WA, Weissman JS, Qi LS. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell*, 2013, 154(2): 442–451. [DOI]
- [84] Perez-Pinera P, Kocak DD, Vockley CM, Adler AF, Kabadi AM, Polstein LR, Thakore PI, Glass KA, Ousterout DG, Leong KW, Guilak F, Crawford GE, Reddy TE, Gersbach CA. RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nat Methods*, 2013, 10(10): 973–976. [DOI]
- [85] Chuai GH, Wang QL, Liu Q. In silico meets *in vivo*: towards computational CRISPR-Based sgRNA design. *Trends Biotechnol*, 2017, 35(1): 12–21. [DOI]
- [86] Echeverri CJ, Beachy PA, Baum B, Boutros M, Buchholz F, Chanda SK, Downward J, Ellenberg J, Fraser AG, Hacohen N, Hahn WC, Jackson AL, Kiger A, Linsley PS, Lum L, Ma Y, Mathey-Prevot B, Root DE, Sabatini DM, Taipale J, Perrimon N, Bernards R. Minimizing the risk of reporting false positives in large-scale RNAi screens. *Nat Methods*, 2006, 3(10): 777–779. [DOI]
- [87] Hart T, Brown KR, Sircoulomb F, Rottapel R, Moffat J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol Syst Biol*, 2014, 10: 733. [DOI]
- [88] Morgens DW, Wainberg M, Boyle EA, Ursu O, Araya CL, Tsui CK, Haney MS, Hess GT, Han K, Jeng EE, Li A, Snyder MP, Greenleaf WJ, Kundaje A, Bassik MC. Genome-scale measurement of off-target activity using Cas9 toxicity in high-throughput screens. *Nat Commun*, 2017, 8: 15178. [DOI]
- [89] Yan JF, Chuai GH, Zhou C, Zhu CY, Yang J, Zhang C, Gu F, Xu H, Wei J, Liu Q. Benchmarking CRISPR on-target sgRNA design. *Brief Bioinform*, 2018, 19(4): 721–724. [DOI]
- [90] Li W, Xu H, Xiao T, Cong L, Love MI, Zhang F, Irizarry RA, Liu JS, Brown M, Liu XS. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol*,

- 2014, 15(12): 554. [DOI]
- [91] Knight SC, Xie L, Deng W, Guglielmi B, Witkowsky LB, Bosanac L, Zhang ET, El Beheiry M, Masson JB, Dahan M, Liu Z, Doudna JA, Tjian R. Dynamics of CRISPR-Cas9 genome interrogation in living cells. *Science*, 2015, 350(6262): 823–826. [DOI]
- [92] Shi J, Wang E, Milazzo JP, Wang Z, Kinney JB, Vakoc CR. Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat Biotechnol*, 2015, 33(6): 661–667. [DOI]
- [93] Hart T, Chandrashekar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmermann M, Fradet-Turcotte A, Sun S, Mero P, Dirks P, Sidhu S, Roth FP, Rissland OS, Durocher D, Angers S, Moffat J. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell*, 2015, 163(6): 1515–1526. [DOI]
- [94] Ong SH, Li Y, Koike-Yusa H, Yusa K. Optimised metrics for CRISPR-KO screens with second-generation gRNA libraries. *Sci Rep*, 2017, 7(1): 7384. [DOI]
- [95] Evers B, Jastrzebski K, Heijmans JP, Gernrum W, Beijersbergen RL, Bernards R. CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. *Nat Biotechnol*, 2016, 34(6): 631–633. [DOI]
- [96] Chen B, Gilbert LA, Cimini BA, Schnitzbauer J, Zhang W, Li GW, Park J, Blackburn EH, Weissman JS, Qi LS, Huang B. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell*, 2013, 155(7): 1479–1491. [DOI]
- [97] Liu H, Wei Z, Dominguez A, Li Y, Wang X, Qi LS. CRISPR-ERA: a comprehensive design tool for CRISPR-mediated gene editing, repression and activation. *Bioinformatics*, 2015, 31(22): 3676–3678. [DOI]
- [98] Horlbeck MA, Witkowsky LB, Guglielmi B, Replogle JM, Gilbert LA, Villalta JE, Torigoe SE, Tjian R, Weissman JS. Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *Elife*, 2016, 5, e12677. [DOI]
- [99] Vidigal JA, Ventura A. Rapid and efficient one-step generation of paired gRNA CRISPR-Cas9 libraries. *Nat Commun*, 2015, 6: 8083. [DOI]
- [100] Zetsche B, Heidenreich M, Mohanraju P, Fedorova I, Kneppers J, Degennaro EM, Winblad N, Choudhury SR, Abudayyeh OO, Gootenberg JS, Wu WY, Scott DA, Severinov K, van der Oost J, Zhang F. Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array. *Nat Biotechnol*, 2017, 35(1): 31–34. [DOI]
- [101] Kim D, Bae S, Park J, Kim E, Kim S, Yu HR, Hwang J, Kim JJ, Kim JS. Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat Methods*, 2015, 12(3): 237–243. [DOI]
- [102] Kim D, Kim S, Kim S, Park J, Kim JS. Genome-wide target specificities of CRISPR-Cas9 nucleases revealed by multiplex Digenome-seq. *Genome Res*, 2016, 26(3): 406–415. [DOI]
- [103] Haeussler M, Kai S, Eckert H, Eschstruth A, Mianné J, Renaud JB, Schneidermaunoury S, Shkumatava A, Teboul L, Kent J, Joly JS, Concordet JP. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol*, 2016, 17(1): 148. [DOI]
- [104] Li C, Zong Y, Wang Y, Jin S, Zhang D, Song Q, Zhang R, Gao C. Expanded base editing in rice and wheat using a Cas9-adenosine deaminase fusion. *Genome Biol*, 2018, 19(1): 59. [DOI]
- [105] Wei Y, Zhang XH, Li DL. The “new favorite” of gene editing technology—single base editors. *Hereditas (Beijing)*, 2017, 39(12): 1115–1121.
魏瑜 张晓辉. 李大力. 基因编辑之“新宠”—单碱基基因组编辑系统. *遗传*, 2017, 39(12): 1115–1121. [DOI]
- [106] Hu JH, Miller SM, Geurts MH, Tang W, Chen L, Sun N, Zeina CM, Gao X, Rees HA, Lin Z, Liu DR. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature*, 2018, 556(7699): 57–63. [DOI]
- [107] Cox DBT, Gootenberg JS, Abudayyeh OO, Franklin B, Kellner MJ, Joung J, Zhang F. RNA editing with CRISPR-Cas13. *Science*, 2017, 358(6366): 1019–1027. [DOI]

(责任编辑: 谷峰)