



دانشگاه صنعتی شریف  
دانشکده علوم ریاضی

پایان نامه کارشناسی ارشد  
ریاضی کاربردی

# تحقیق در ویرایش ژنوم به کمک تناوب‌های کوتاه پالیندروم فاصله‌دار منظم خوشه‌ای

نگارش

محمد رستمی

استاد راهنما

دکتر محسن شریفی تبار

استاد راهنمای دوم

دکتر حمیدرضا ربیعی

استاد مشاور

دکتر محمدحسین رهبان

20 بهمن 1400

به نام او  
دانشگاه صنعتی شریف  
دانشکده علوم ریاضی

پایان نامه کارشناسی ارشد  
عنوان: تحقیق در ویرایش ژنوم به کمک تناوب‌های کوتاه پالیندروم فاصله‌دار منظم خوشه‌اینگارش: محمد رستمی

کمیته داوران

امضاء:.....	دکتر محسن شریفی تبار	استاد راهنما:
امضاء:.....	دکتر حمیدرضا ربیعی	استاد راهنمای همکار:
امضاء:.....	دکتر محمدحسین رهبان	استاد مشاور:
امضاء:.....	1	ممتحن داخلی:
امضاء:.....	2	ممتحن داخلی:
امضاء:.....	3	داور خارجی:
امضاء:.....	4	داور خارجی:
تاریخ:.....		

## قدردانی

با تشکر از دکتر ربیعی، دکتر رهبان، استاد راهنمای عزیزم دکتر شریفی تبار، امین قریاضی و حامد دشتی برای کمک‌های مداومشان، و تشکر از آقای وفا خلیقی که با طراحی بسته XqPersian کمک بزرگی به حروف‌چینی فارسی کردند،

و تشکر از خداوند.

### چکیده

تناوب‌های کوتاه پالیندروم فاصله‌دار منظم خوشه‌ای یا به طور خلاصه ، CRISPR (کریسپر) یکی از روش‌های نسبتاً نوین است که متخصصان ژنتیک و محققان پزشکی را قادر می‌سازد تا با حذف بخش‌هایی از ژنوم ، افزودن یا تغییر بخش‌هایی از آن در dna (دی‌ان‌ای) تغییر ایجاد کنند. این فناوری نوعی سیستم ایمنی تطابق‌پذیر در باکتری‌ها است که با کمک آن می‌توان بسیاری از بیماری‌ها مانند نابینوایی و ناشنوایی و حتی سرطان را درمان کرد. یکی از مشکلات بزرگ در استفاده موفق کریسپر، دقیق پیش‌بینی کردن تاثیر Guide RNA (راهنمای آران‌ای) روی هدف و حساسیت خارج از هدف است. در حالی که برخی از روش‌ها این طرح‌ها را طبقه‌بندی می‌کنند ، بیشتر الگوریتم‌ها بر روی داده‌های جداگانه با ژن‌ها و سلول‌های مختلف هستند. عدم تعمیم این روش‌ها مانع استفاده از این راهنما در آزمایشات بالینی می‌شود ، زیرا برای هر درمان، این فرایند باید دقیقاً برای همان سلول درست شده باشد و عموماً داده کافی برای طراحی الگوریتم در آن سلول در دسترس نیست. ما سعی می‌کنیم مشکل تعمیم‌پذیری را حل کنیم و مدل‌ای ارائه دهیم که هم به صورت عمومی و هدفمند دقت خوبی داشته باشد تا که محققان در بهینه‌سازی طراح راهنمای آران‌ای با حساسیت مناسب کمک کند.

# فهرست مطالب

1	مقدمات	1	
1	RNA	1.1	
1	دی‌ان‌ای	2.1	
2	تفاوت‌های دی‌ان‌ای و آران‌ای	1.2.1	
3	ویرایش ژنوم	3.1	
3	شکست و تعمیر دی‌ان‌ای	1.3.1	
4	Zinc finger nucleases (ZFN)	2.3.1	
4	TALEN	3.3.1	
5	CRISPR	4.1	
5	کریسپر در باکتری	1.4.1	
5	عمل کرد کریسپر در ژن	2.4.1	
6	حساسیت	3.4.1	
6	تاثیرگذاری	4.4.1	
7	انواع کریسپر	5.4.1	
10	Literature the of Review	2	
10	روش‌های مستقیم	1.2	
10	[29-27] Chopchop	1.1.2	
10	[26, 25] Cas-Designer و Cas-OFFinder	2.1.2	
11	[24] E-CRISP	3.1.2	
12	[23] CRISPOR	4.1.2	
12	روش‌های یادگیری ژرف	2.2	
12	پیش‌بینی off-target به کمک یادگیری ژرف	1.2.2	
13	[19] CCTop	2.2.2	
13	[22] DeepCRISPR	3.2	
14	Methods	3	
14	ensemble	1.3	
14	تعریف	1.1.3	
15	Attention	2.3	
16	Results	4	
16	Ensemble	1.0.4	
19	Attention	2.0.4	
21	Discussion	5	

# فهرست تصاویر

1	1.1	یک حلقه از pre-mRNA. نوکلئوبازها (سبز) و ستون فقرات ریبوز فسفات (آبی) مشخص شده اند. این یک رشته منفرد از RNA است که بر روی خود تا می شود.
2	2.1	شکل دو بعدی DNA
2	3.1	مقایسه دیانای و آرانای
3	4.1	مکانیزم ترمیم DNA
4	5.1	مکانیزم TALEN
5	6.1	مکانیزم ساده شده ای از CRISPR
6	7.1	مکانیزم CRISPR
7	8.1	مکانیزم TALEN
11	1.2	(الف) شماتیک مکان های off-targets را با برآمدگی DNA یا RNA نشان می دهد. (ب) استراتژی برای برآمدگی 1-nt DNA یا RNA بر اساس Cas-OFFinder. (ج) یک مثال از یک جدول خروجی Cas-Designer تمام gRNA های ممکن را از توالی های ورودی به همراه اطلاعات مفید (بالا) نشان می دهد. اگر کاربر روی رنگ آبی کلیک کند عدد، کلمه یا عبارت، اطلاعات دقیق تری مانند اهداف برآمدگی DNA (وسط) یا RNA (پایین) ارائه می شود. علاوه بر این، کاربر می تواند موارد مربوطه را به دست آورد اطلاعات ژنومی از طریق مرورگر ژنوم Ensembl (Flicek و همکاران، 2011)، با کلیک بر روی دکمه "اطلاعات در Ensembl"
11	2.2	الگوریتم E-CRISP
12	3.2	هم بستگی اسپیرمن بین امتیاز موثر بودن و داده ها
12	4.2	هم بستگی اسپیرمن بین امتیاز موثر بودن و داده ها [21]
13	5.2	DeepCRISPR
16	1.4	ROC AUC
16	2.4	PR AUC
17	3.4	score F1
17	4.4	هم بستگی اسپیرمن بین امتیاز موثر بودن و داده ها
18	5.4	نتیجه آموزش ل
18	6.4	هم بستگی اسپیرمن بین امتیاز موثر بودن و داده ها
19	7.4	نتیجه تمرین به کمک 3mer، به کمک مدل DNAbert
19	8.4	نتیجه تمرین به کمک 4mer، به کمک مدل DNAbert
19	9.4	نتیجه تمرین به کمک 6mer، به کمک مدل DNAbert
20	10.4	نتیجه تمرین به کمک 6mer، به کمک مدل RoBerta

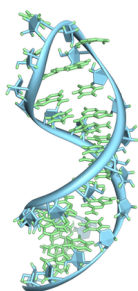
# فصل 1

## مقدمات

قبل از این که با کریسپر آشنا شویم، خوب است کمی راجع به تاریخچه ویرایش ژن ها صحبت کنیم. انسان ها سال ها است که مشغول به ویرایش و مهندسی ژن هستند، با استفاده از پرورش انتخابی<sup>1</sup>. اصلاحات نژادی متعددی در گیاهان و حیوانات مخصوصا گونه های کلیدی مانند گندم، برنج و سگ ها ایجاد شده است. انسان ها در این کار شدیداً ماهر شدن به طوری که در صده گذشته، تعداد دانه های هر شاخه گندم چندین برابر و ارتفاع آنها کوتاه تر شده تا در معرض خطر کمتری باشند و حدود ۸۰ نژاد جدید سگ به وجود آمده است. البته با وجود پیشرفت های متعدد انسان ها تا کشف دی ان ای دقیقاً ساز و کار آن را نمی دانستند.

### 1.1 RNA

اسید ریبونوکلیک<sup>2</sup> یا RNA یک مولکول پلیمری است که در نقش های بیولوژیکی مختلف مانند کدگذاری، رمزگشایی، تنظیم و بیان ژن ها ضروری است. RNA به صورت یک رشته منفرد از نوکلئوتیدها (بازهای نیتروژنی گوانین، اوراسیل، آدنین و سیتوزین که با حروف G، A، U و C مشخص می شوند) است که برخوردش تا می خورد، بر خلاف دی ان ای که با یک رشته دیگر جفت شده است.



نوعی از RNA اطلاعات را از DNA به سیتوپلاسم حمل می کند؛ به این نوع RNA که اطلاعات را از DNA به ریبوزوم ها حمل می کند، RNA پیک یا پیامبر (mRNA) می گویند. نوعی دیگر از RNA، RNA حامل (tRNA) است که اسیدهای آمینه را به ریبوزوم منتقل می کند، تا ریبوزوم، اسیدهای آمینه را بر اساس اطلاعات موجود در mRNA کنار یکدیگر ردیف کند. نوع دیگر، RNA ریبوزومی (rRNA) است که در ساختار ریبوزوم ها شرکت دارد؛ این موضوع به این معناست که ریبوزوم (رئاتن) ها متشکل از پروتئین ها و RNA های ریبوزومی هستند.

شکل 1.1: یک حلقه از pre-mRNA نوکلئوبازها (سبز) و ستون فقرات ریبوز فسفات (آبی) مشخص شده اند. این یک رشته منفرد از RNA است که بر روی خود تا می شود.

### 2.1 دی ان ای

دئوکسی ریبو نوکلئیک اسید<sup>3</sup> به اختصار دی ان ای یک مولکول متشکل از دو زنجیره پلی نوکلئوتیدی است که به دور یکدیگر می پیچند تا دستورالعمل های ژنتیکی برای کارکرد و توسعه زیستی جانداران و ویروس ها مورد استفاده قرار می گیرد. نقش اصلی مولکول دی ان ای ذخیره سازی طولانی مدت اطلاعات ژنتیکی و دستوری است. لیپیدها، پروتئین ها، کربوهیدرات های پیچیده (پلی ساکاریدها) و اسیدهای نوکلئیک سه درشت مولکول های اصلی و ضروری برای همه اشکال شناخته شده حیات هستند.

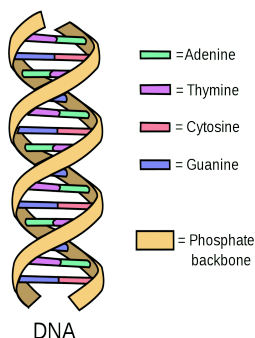
دو رشته DNA به عنوان پلی نوکلئوتید شناخته می شوند زیرا از واحدهای مونومر یا تکپار ساده تری به نام نوکلئوتید تشکیل شده اند. هر نوکلئوتید از یکی از چهار نوکلئوباز حاوی نیتروژن (سیتوزین، گوانین، آدنین A یا تیمین، T) کربوهیدرات پنج کربنه به نام دئوکسی

<sup>1</sup>Selective Breeding

<sup>2</sup>RiboNucleic Acid

<sup>3</sup>Deoxyribonucleic acid

ریبوز و یک گروه فسفات تشکیل شده است. نوکلئوتیدها در یک زنجیره توسط پیوندهای کووالانسی (معروف به پیوند فسفو دی استر) بین قند یک نوکلئوتید و فسفات نوکلئوتید بعدی به یکدیگر متصل می‌شوند و در نتیجه یک ستون فقرات قند-فسفات متناوب ایجاد می‌شود. بازهای نیتروژنی دو رشته پلی نوکلئوتیدی جداگانه، طبق قوانین جفت شدن بازها (A با T و C با G)، با پیوندهای هیدروژنی به یکدیگر متصل می‌شوند تا DNA دو رشته‌ای بسازند. این دو رشته مکمل، ناهمسو و محلول (در آب) هستند (دی‌ان‌ای حلقوی قطبیت ندارد اما هر رشته از دی‌ان‌ای خطی دارای قطبیت است). بازهای نیتروژنی مکمل به دو گروه پیریمیدین‌ها و پورین‌ها تقسیم می‌شوند. در DNA، پیریمیدین‌ها تیمین و سیتوزین هستند. پورین‌ها آدنین و گوانین هستند.



شکل 2.1: شکل دو بعدی DNA

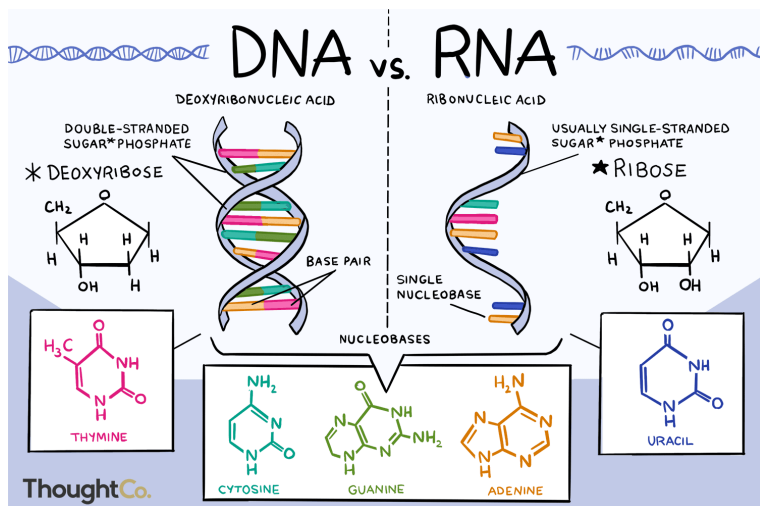
هر دو رشته DNA اطلاعات بیولوژیکی یکسانی را ذخیره می‌کنند. این اطلاعات زمانی که دو رشته از هم جدا می‌شوند، تکرار می‌شود. بخش بزرگی از DNA (بیش از 98٪ برای انسان) بی‌کد<sup>4</sup> است، به این معنی که این بخش‌ها توالی‌های پروتئین را کد نمی‌کنند. دو رشته DNA در جهت مخالف یکدیگر قرار دارند و بنابراین باز مکمل ابتدای یک رشته آخر رشته دیگر هستند. در آیین نامگذاری ترکیبهای شیمیایی، اتمهای کربن در حلقه شکر نوکلئوتید شماره‌گذاری شده‌اند. هر رشته دی‌ان‌ای یا آر‌ان‌ای دارای یک پایانه 5' که معمولاً شامل یک گروه فسفاتی است و یک پایانه 3' که معمولاً از جانشین ریبوز اصلاح نشده OH- است. به هر قند یکی از چهار نوع نوکلئوباز (یا باز) متصل است. توالی این چهار هسته در امتداد ستون فقرات است که اطلاعات ژنتیکی را رمزگذاری می‌کند. رشته‌های RNA با استفاده از رشته‌های DNA به‌عنوان یک الگو در فرآیندی به نام رونویسی ایجاد می‌شوند که در آن بازهای DNA با بازهای مربوطه خود مبادله می‌شوند، به جز در مورد تیمین (T) که RNA جایگزین اوراسیل (U) می‌شود. تحت کد ژنتیکی، این رشته‌های RNA توالی اسیدهای آمینه درون پروتئین‌ها را در فرآیندی به نام ترجمه مشخص می‌کنند.

### 1.2.1 تفاوت‌های دی‌ان‌ای و آر‌ان‌ای

تفاوت‌ها:

- DNA در ذخیره و RNA در انتقال اطلاعات وراثتی و در ساختار ریبوزوم نقش دارد.
- مولکول DNA دو رشته‌ای در هم تنیده اما مولکول RNA تک‌رشته‌ای است.
- در DNA باز آلای یوراسیل و در RNA باز آلای تیمین شرکت ندارد (U در DNA و T در RNA).
- قند پنج کربنه موجود در DNA را دئوکسی ریبوز و در RNA قند ریبوز نامیده می‌شود. تفاوت بین قندها وجود گروه هیدروکسیل بر روی کربن 2' ریبوز و عدم وجود آن در کربن 2' دئوکسی ریبوز است.
- DNA برعکس RNA از هسته سلول خارج نمی‌شود.
- RNA بدون ژن می‌باشد.

شباهت‌ها:



- هر دو پلیمر هستند و از نوکلئوتید تشکیل شده‌اند.
- در هر دو نوکلئوتیدهای مقابل با پیوند هیدروژنی و نوکلئوتیدهای کناری با پیوند فسفو دی‌استر به هم متصل می‌شوند (گاهی نوکلئوتیدهای دو بخش متفاوت از یک رشته آر‌ان‌ای، به هم متصل می‌شوند).
- نوکلئوتیدهای آزاد (واحدهای سازنده آزاد) هر دو مولکول پیش از اتصال سه فسفات بوده و با اتصال به رشته پلی‌نوکلئوتیدی تک‌فسفاته می‌شوند.

شکل 3.1: مقایسه دی‌ان‌ای و آر‌ان‌ای <sup>4</sup>non-coding



### 3.1 ویرایش ژنوم

مهندسی ژنوم یا ویرایش ژنوم نوعی از مهندسی ژنتیک است که در آن دی‌ان‌ای ژنوم یک موجود زنده حذف، اضافه، اصلاح یا جایگزین می‌شود. در دهه ۱۹۶۰، دانشمندان با شارش پرتوهای رادیواکتیوی بر روی گیاهان دست به تغییر ژنوم آنها به طور کاملاً تصادفی زدند. این کار به هدف رسیدن به یک تغییر ژنتیک مفید صورت می‌گرفت و البته نتایج خوبی هم به همراه داشت ولی با این حال راندمان پایین این ویرایش‌ها باعث شد که دانشمندان به فکر راه‌های دیگری برای ویرایش ژنوم باشند.

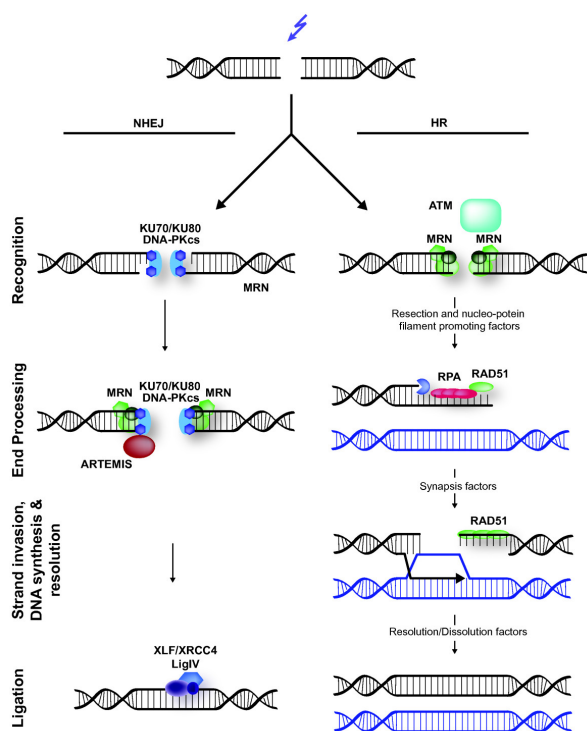
تا کنون سه تکنیک موفق و معروف برای ویرایش ژنوم مهندسی شده است: نوکلئاز انگشت روی<sup>۵</sup> (ZFNs)، نوکلئازهای اثرگذار شبه فعال کننده رونویسی<sup>۶</sup> (TALEN)، و سیستم تناوب‌های کوتاه پالیندروم فاصله‌دار منظم خوشه‌ای<sup>۷</sup> (CRISPR). کلید ویرایش ژنوم ایجاد شکست دو رشته‌ای DNA در نقطه مورد نظر است و این سه روش مبتنی بر شکست درست DNA در نقطه مورد نظر مهندسی شدند.

#### 1.3.1 شکست و تعمیر دی‌ان‌ای

شکل رایج ویرایش ژنوم بر مفهوم مکانیک ترمیم شکست دو رشته‌ای DNA<sup>۸</sup> (DSB) تکیه دارد. دو مسیر اصلی وجود دارد که DSB را تعمیر می‌کند. اتصال انتهای غیر همولوگ<sup>۹</sup> (NHEJ) و تعمیر هدایت شده همولوژی<sup>۱۰</sup> (HDR). NHEJ از انواع آنزیم‌ها برای اتصال مستقیم به انتهای DNA استفاده می‌کند، در حالی که HDR دقیق‌تر از یک توالی همولوگ به عنوان الگویی برای بازسازی توالی‌های DNA گمشده در نقطه شکست استفاده می‌کند. این را می‌توان با ایجاد یک بردار با عناصر ژنتیکی مورد نظر در یک توالی که همولوگ با توالی‌های کناری یک DSB است مورد استفاده قرار داد. این باعث می‌شود که تغییر مورد نظر در محل DSB درج شود. در حالی که ویرایش ژن مبتنی بر HDR مشابه هدف‌گیری ژن مبتنی بر نوترکیب همولوگ است، نرخ نوترکیبی حداقل سه مرتبه افزایش می‌یابد.

##### NHEJ

پرتوهای یونیزه کننده و برخی داروهای ضد سرطان باعث شکست هر دو رشته‌ای DNA می‌شوند. سیستمی که برای ترمیم این نوع آسیب به کار گرفته می‌شود، سیستم ترمیم اتصال انتهای غیر همولوگ (NHEJ) می‌باشد که مستعد به خطا به شمار می‌رود، زیرا همواره چندین نوکلئوتید در جایگاه ترمیم از بین می‌روند و دو انتهای شکسته شده از کروموزوم‌های همولوگ یا غیر همولوگ به یکدیگر متصل می‌شوند. زمانی که کروماتیدهای خواهری برای ترمیم شکست‌های دو رشته‌ای در دسترس نباشند این نوع ترمیم صورت می‌گیرد. در ابتدا کمپلکسی از ku70/80 و پروتئین کیناز وابسته به DNA به انتهاهای شکسته دو رشته اتصال می‌یابند، آن‌گاه در هر انتها چندین باز توسط نوکلئاز حذف شده و دو مولکول از طریق آنزیم لیگاز به هم متصل می‌گردند. DSB‌ها ترجیحاً در سلول توسط اتصال انتهای غیر همولوگ (NHEJ) ترمیم می‌شوند، مکانیزم سریعی که اغلب باعث درج یا حذف (indels) در DNA می‌شود. ایندل‌ها اغلب منجر به تغییر اساسی در DNA می‌شوند و به‌طوری که DNA عملکرد خود را از دست می‌دهند. پس در نتیجه معمولاً به عنوان سلول مرده در نظر گرفته می‌شوند و حذف می‌شوند. برای ویرایش ژنوم مطمئن اعمال شدن ویرایش و تغییر نکردن آن نکته مهمی است. پس دانشمندان تمام تلاششان را میکنند که بعد از DNA DSB، به این روش تعمیر نشود.



شکل 4.1: مکانیزم ترمیم DNA

<sup>۵</sup> Zinc Finger Nuclease

<sup>۶</sup> Transcription activator-like effector nuclease

<sup>۷</sup> Clustered Regularly Interspaced Short Palindromic Repeats

<sup>۸</sup> Double-Strand Break (Cut)

<sup>۹</sup> Non-Homologous End Joining

<sup>۱۰</sup> Homology Directed Repair

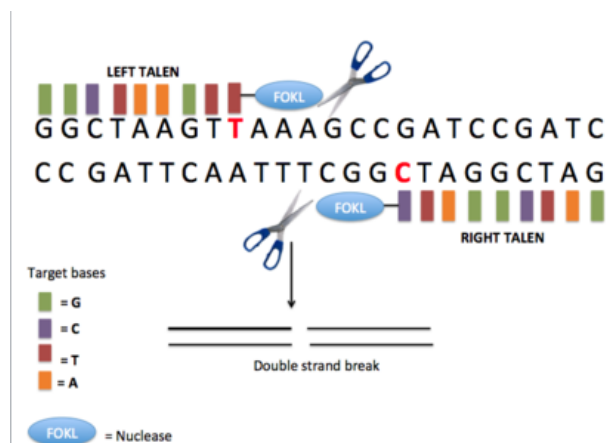
## HDR

تعمیر هدایت شده همولوژی (HDR) مکانیزمی در سلول‌ها برای ترمیم ضایعات DNA دو رشته‌ای از یک نسخه مشابه DNA برای ترمیم استفاده می‌کند. رایج‌ترین شکل HDR نوترکیبی همولوگ است. مکانیسم HDR تنها زمانی می‌تواند توسط سلول استفاده شود که یک قطعه همولوگ از DNA در هسته وجود داشته باشد، عمدتاً در فاز G2 و S چرخه سلولی. نمونه‌های دیگر تعمیر مبتنی بر HDR شامل تعمیر تک رشته‌ای و تکثیر ناشی از شکستگی است. هنگامی که DNA همولوگ وجود ندارد، فرآیند NHEJ به جای آن انجام می‌شود.

## Zinc finger nucleases (ZFN) 2.3.1

نوکلئاز انگشت روی یا ZFN اولین سیستم پروتئینی متصل شونده به DNA قابل برنامه‌ریزی با کاربرد وسیع است. ZFNها شامل زنجیره‌ای از پروتئین‌های انگشت روی هستند که به یک نوکلئاز باکتریایی ملحق شده‌اند تا بتوانند سیستمی را تولید کنند که قادر به ایجاد برش‌های دو رشته‌ای خاص در DNA برای ویرایش ژن باشد. پروتئین‌های انگشت روی هدف قرار دادن ناحیه خاص را فراهم می‌کنند زیرا هر یک از آنها سه جفت باز یا ۳bp از DNA را شناسایی می‌کنند. نوکلئازی که معمولاً در تکنولوژی ZFN متصل به زنجیره پروتئین‌های انگشت روی است FokI نام دارد که برای اتصال به DNA باید دایمریزه شده باشد، بنابراین یک جفت از ZFN برای هدف‌گیری و برش DNA مورد استفاده قرار می‌گیرد. این آنزیم‌ها کمک زیادی به تولید موجودات ترانسژنیک می‌کنند و بدلیل اینکه فراوانی نوترکیبی همولوگ بسیار ناچیز بوده اهمیت زیادی در مهندسی ژنتیک و مطالعات ترانسژنیک، ناک‌اوت و غیره پیدا کرده‌اند. این پروتئین‌های مهندسی شده متصل شونده به DNA می‌توانند ژنوم را در جایگاه‌های ویژه‌ای شناسایی کرده و ایجاد برش‌های دورشته‌ای کنند. در صورتیکه سیستم تعمیر NHEJ فعال شود چون این سیستم ترمیم مستعد خطاست سبب ایجاد جهش در آن ناحیه خاص از ژنوم می‌شود بنابراین در مطالعات موتاژن نیز اهمیت دارند. انتقال یک وکتور حاوی ژن مورد نظر به همراه ZFNs سبب تسهیل درج ژن در آن ناحیه از ژنوم می‌گردد.

## TALEN 3.3.1



شکل 5.1: مکانیزم TALEN

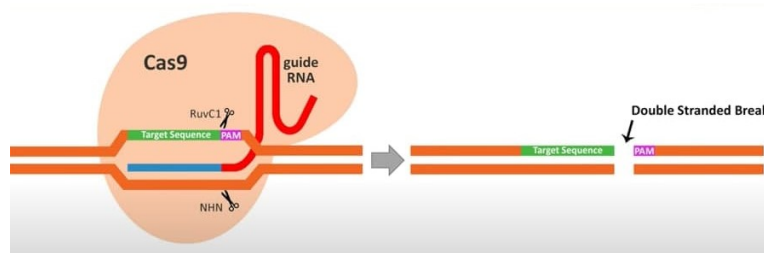
نوکلئازهای رونویس مؤثر-مانند فعال‌کننده (TALENs) پروتئین‌های خاصی هستند که به DNA متصل می‌شوند که دارای آرایه‌ای از 33 یا 34 تکرار آمینه اسیدها هستند. TALEN ها آنزیم‌های محدودکننده مصنوعی هستند که با ادغام حوزه برش DNA یک نوکلئاز با دامنه‌های TALE طراحی شده‌اند، که می‌توانند به طور خاص یک توالی DNA منحصر به فرد را شناسایی کنند. این پروتئین‌های ادغام شده به عنوان قیچی DNA به راحتی قابل برنامه‌نویسی برای ویرایش یک ژن خاص عمل می‌کنند که قادر به انجام تغییرات هدفمند ژنوم مانند درج توالی، حذف، تعمیر و جایگزینی در سلول‌های زنده هستند. این تکنولوژی را می‌توان برای تغییر هر نقطه از DNA استفاده کرد. TALهای مؤثر یک رشته ۳۴ تایی از آمینواسیدها هستند که هر کدام وظیفه دارند یک تک نوکلئوتید را پیدا کنند. نوکلئاز می‌تواند شکستگی‌های دو رشته‌ای را در محل هدف ایجاد کند که می‌تواند با اتصال انتهایی غیر همولوگ (NHEJ) ترمیم شود، که منجر به اختلالات ژنی از طریق وارد کردن یا حذف‌های کوچک می‌شود. هر تکرار حفظ می‌شود، به جز آمینواسید ۱۲ و ۱۳ که به آنها دو باقیمانده متغیر تکرار (RVDs) می‌گویند.

ها توالی DNA را تعیین می‌کنند که TALE به آن متصل می‌شود. این تناظر ساده یک به یک بین تکرارهای TALE و توالی DNA مربوطه باعث می‌شود روند مونتاژ آرایه‌های تکراری برای تشخیص توالی‌های DNA جدید ساده باشد. این TALEها را می‌توان با کاتالیزوری از یک نوکلئاز از DNA به نام FokI، ادغام کرد تا با آن‌ها TALEN را ساخت. ساختارهای TALEN توالی‌های DNA را فقط در مکان‌های از پیش انتخاب شده متصل می‌کنند و می‌شکنند. هدف TALEN را می‌توان بر اساس یک کد آسان پیش‌بینی کرد. نوکلئازهای TAL تا حدی به دلیل طولشان که بیش از 30 جفت است می‌توان فقط مختص آن هدف در نظر گرفت. TALEN را می‌توان در محدوده 6 جفت باز از هر نوکلئوتید منفرد در کل ژنوم انجام داد.

سازه‌های TALEN به روشی مشابه با نوکلئازهای انگشت روی طراحی شده استفاده می‌شوند و دارای سه مزیت در جهش‌زایی هدفمند هستند: (1) اختصاصیت اتصال به DNA بالاتر است، (2) اثرات خارج از هدف کمتر است و (3) طراحی آن آسان‌تر است.

## 4.1 CRISPR

CRISPR (به انگلیسی: Clustered Regularly Interspaced Short Palindromic Repeats) یا به اختصار کریسپر (به معنی "تناوب‌های کوتاه پالیندروم فاصله‌دار منظم خوشه‌ای" بخشی از دی‌ان‌ای پروکاریوت هستند که حاوی تناوب‌های کوتاه توالی‌های بنیادین هستند. بخشی از سیستم کریسپر "پروتئین Cas9" است. این پروتئین قابلیت جستجو، برش زدن و تغییر دی‌ان‌ای (DNA) را دارد. قبل از این تکنیک از روش "تحویل یا انتقال ژن" استفاده می‌شد، به این صورت که از یک ناقل ویروسی یا غیروبیروسی برای انتقال ژن سالم به ژنوم سلول میزبان استفاده می‌شد، ولی در روش کریسپر، ژن معیوب برش داده می‌شود و ژن سالم به جای آن قرار می‌گیرد. استفاده از آنزیم Cas9 خطر کمتری نسبت به روش قبلی که یک ژن خارجی وارد ژنوم می‌شد دارد، زیرا گاهی ژن خارجی به سرطان منجر می‌شود اما ژنی که از طریق کریسپر ترمیم شود کنترل شده است. نام دیگر این تکنیک "قیچی ژنتیکی" است که به دلیل ساز و کار آنزیم "کس9" (Cas9) هست. این آنزیم به عنوان یک جفت قیچی مولکولی می‌تواند دو رشته DNA را در محل خاصی از ژنوم برش دهد. [۱]



شکل 6.1: مکانیزم ساده شده‌ای از CRISPR

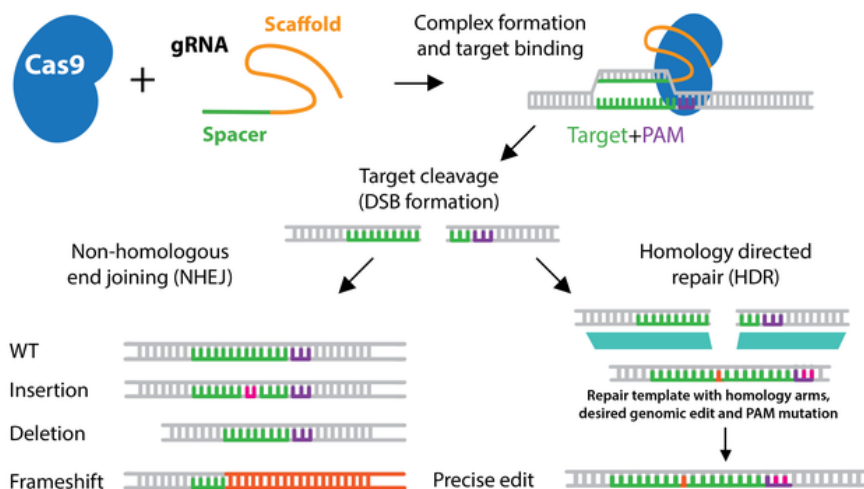
## 1.4.1.1 کریسپر در باکتری

اولین بار سیستم کریسپر در *Escherichia coli* به عنوان یک توالی تکراری ۲۹ نوکلئوتیدی با فاصله ۳۲ نوکلئوتیدی توسط یوشیزومی ایشی نو ژاپنی در سال ۱۹۸۷ مطرح شد که باکتری‌ها و آرکی باکتری‌ها را از حمله باکتریوفاژها و پلاسمیدها محافظت می‌کند. این سیستم‌های دفاعی به یک RNA کوچک شناساگر توالی خاص تکیه می‌کنند و اسیدهای نوکلئیک خارجی را خاموش می‌کنند. [۲] Francisco Mojica و همکارانش در سال ۱۹۹۳ تکرارهای مشابهی را در چندین گونه میکروبی دیگر یافتند. [۳] بعد از حمله به سلول توسط عناصر ژنتیکی خارجی مانند باکتریوفاژها یا پلاسمیدها (مرحله ۱: تزریق فاژ)، آنزیم‌های ویژه مرتبط CRISPR به نام Cas (CRISPR-associated protein) توالی‌های spacer را از توالی‌های protospacer جدا کرده و آن‌ها را به درون لوکوس‌های کریسپر موجود در ژنوم پروکاریوت‌ها وارد و متصل می‌کنند. (مرحله ۲: استفاده از spacer این spacerها بین تکرارهای مستقیم تقسیم شده‌اند که اجازه می‌دهند سیستم CRISPR به‌طور ایمن و دقیق و نه به‌طور غیر ایمن شناسایی شود. آرایه CRISPR یک رونوشت RNA غیر کدونی است که از نظر آنزیمی از طریق مسیرهای متمایز که برای هر نوع سیستم CRISPR منحصر به فرد است، بالغ می‌شود. (مرحله ۳: بیوژنز و پردازش CrRNA در CRISPR نوع I و III، رونوشت pre-CrRNA توسط ریبونوکلازهای مرتبط با CRISPR، شکسته می‌شوند و این کار موجب آزاد شدن چندین CrRNAs کوچک می‌شود. به‌طور متوسط CrRNA نوع III بیشتر در انتهای ۳' توسط RNase‌های که هنوز مشخص نشده‌اند برای تولید رونوشت کاملاً بالغ پردازش می‌شوند. CRISPR نوع II، یک کریسپر فعال کننده ترانس است (tracrRNA) که با تکرارهای مستقیم هیبرید می‌شود و یک RNA دوپلکس را تشکیل می‌دهد و توسط RNase III درونی و نوکلئازهای ناشناخته دیگر شکسته و پردازش می‌شود. های CrRNA بالغ شده نوع I و III سیستم CRISPR، سپس درون افکتورهای کمپلکس‌های پروتئینی برای تشخیص و تخریب توالی هدف اضافه می‌شوند. در سیستم‌های نوع II، کمپلکس هیبرید CrRNA-tracrRNA به Cas9 متصل شده و در واقع هیبرید شدن این دو باعث فعال شدن Cas9 می‌شود. هر دو نوع I و III سیستم CRISPR از چند پروتئین مداخله گر تنظیم‌کننده برای تسهیل شناسایی توالی هدف استفاده می‌کنند. در CRISPR نوع I، کمپلکس Cascade با یک مولکول CrRNA لود می‌شود که یک مجموعه نظارتی بی نظیری است که DNA هدف را شناسایی می‌کند. سپس نوکلئاز Cas3، لوپ Cascade R را به کار گرفته و به آن متصل می‌شود و واسطه تخریب توالی هدف می‌شود. در CRISPR نوع III، CrRNA‌ها یا به کمپلکس‌های Csm یا به کمپلکس‌های Cmr به ترتیب متصل شده و به ترتیب سوبستراهای DNA و RNA را می‌شکنند. در مقابل، سیستم نوع II فقط نیاز به Cas9 برای تخریب DNA جفت شده با RNA راهنما دوپلکس خود دارد که این RNA راهنما حاوی ترکیبی از CrRNA-tracrRNA است. [۵]

## 2.4.1.1 عمل کرد کریسپر در ژن

همانطور که گفتیم، مدل‌های مختلفی از CRISPR تا به حال درست شده ولی به صورت کلی می‌توان CRISPR را به دو قسمت RNA و Cas تقسیم کرد که Cas در آن وظیفه جدا کردن دو رشته DNA را از هم دارد و RNA که هدف را مشخص و قیچی می‌کند. برای این که دقیقاً نقطه شکست DNA مشخص شود Cas نیاز به یک سیگنال است که با رسیدن به آن کار خود را شروع کند. به این رشته PAM یا Protospacer Adjacent Motif گفته می‌شود که کاملاً وابسته به Cas است. همان طور که از قبل گفتیم بعد از شکست

DNA، دو مکانیزم برای تعمیر آن وجود دارد. دانشمندان تکنولوژی‌های CRISPR مختلفی را برای افزایش احتمال تعمیر HDR ایجاد کرده‌اند که با هر یک ویژگی‌های خاص خود را دارند ولی ما در پژوهش خود ساده‌ترین مورد آن یعنی Cas9 به همراه یک RNA که به آن single guide RNA یا sgRNA می‌گویند، استفاده کرده‌ایم. این طرح باعث محدود شدن هدف‌های مورد استفاده می‌شود به طوری که PAM باید به شکل NGG باشد که در آن N یک نوکلئوتید دلخواه است. در نتیجه رشته هدف همیشه با NGG ختم می‌شود.



شکل 7.1: مکانیزم CRISPR

### 3.4.1 حساسیت

حساسیت در یک طرح CRISPR میزان اختصاصی بودن توالی هدف‌گیری شده توسط gRNA در مقایسه با بقیه ژنوم تعیین می‌شود. در حالت ایده‌آل، یک توالی هدف‌گیری شده توسط gRNA همسانی کاملی با DNA هدف خواهد داشت و هیچ همسانی در جای دیگری در ژنوم وجود ندارد یعنی دقیقاً هدف را ویرایش می‌دهد نه جای دیگری را. با این حال، به طور واقع بینانه، یک توالی که با gRNA هدف قرار گرفته شده، مکان‌های بیشتری در سراسر ژنوم ویرایش خواهد داد که در آن همولوژی نسبی وجود دارد. این ناحیه‌ها خارج از هدف یا offtarget نامیده می‌شوند و باید هنگام طراحی یک gRNA برای آزمایش خود در نظر گرفته شوند.

علاوه بر بهینه‌سازی طراحی، gRNA حساسیت CRISPR نیز می‌تواند از طریق تغییرات در Cas9 افزایش یابد. همانطور که قبلاً بحث شد، Cas9 از طریق فعالیت ترکیبی دو حوزه نوکلئاز، HNH و RuvC، شکست‌های دو رشته‌ای (DSBs) ایجاد می‌کند. نیکاز Cas9 یک جهش D10A از SpCas9، یک دامنه نوکلئاز را حفظ می‌کند و به جای DSB، یک DNA نیک تولید می‌کند.

بنابراین، دو نیکاز که رشته‌های DNA مخالف را هدف قرار می‌دهند برای تولید DSB در DNA هدف مورد نیاز است. این نیاز برای یک سیستم CRISPR نیکاز دوتایی یا نیکاز دوگانه به طور چشمگیری ویژگی هدف را افزایش می‌دهد، زیرا بعید است که دو ناک خارج از هدف به اندازه کافی نزدیک به ایجاد DSB ایجاد شوند. اگر حساسیت بالا برای آزمایش شما بسیار مهم است، ممکن است استفاده از رویکرد نیکاز دوگانه را برای ایجاد یک DSB القا شده با نیک دوگانه در نظر بگیرید. سیستم نیکاز همچنین می‌تواند با ویرایش ژن با واسطه HDR برای ویرایش‌های ژنی خاص ترکیب شود.

در سال 2015، محققان از rational mutagenesis برای توسعه دو Cas9 با ثبات بالا استفاده کردند: eSpCas9 و SpCas9-HF1. SpCas9 حاوی جایگزین‌های آلانین است که برهمکنش‌های بین شیار HNH/RuvC و رشته DNA غیرهدف را تضعیف می‌کند و از جدا شدن رشته‌ها و برش در مکان‌های خارج از هدف جلوگیری می‌کند. به طور مشابه، SpCas9-HF1 ویرایش خارج از هدف را از طریق جایگزینی آلانین کاهش می‌دهد که برهمکنش Cas9 با ستون فقرات فسفات DNA را مختل می‌کند. یکی دیگر از Cas9 با وفاداری بالا، HypaCas9، در سال 2017 توسعه یافت و حاوی جهش‌هایی در دامنه REC3 است که تصحیح Cas9 و تبعیض هدف را افزایش می‌دهد. هر سه آنزیم با وفاداری بالا نسبت به Cas9 نوع وحشی ویرایش خارج از هدف کمتری تولید می‌کنند.

### 4.4.1 تاثیرگذاری

تاثیرگذاری در یک طرح CRISPR احتمال شکست DNA و ویرایش درست را تعیین می‌کند. برای غلبه بر راندمان پایین HDR، محققان دو دسته از ویرایشگرهای پایه را ایجاد کرده‌اند - ویرایشگرهای پایه سیتوزینی (CBEs) و ویرایشگرهای پایه آدنین (ABEs).

ویرایشگرهای پایه سیتوزینی با ادغام نیکاز Cas9 یا Cas9 مرده غیرفعال کاتالیزوری (dCas9) به سیتیدین دامیناز مانند APOBEC ایجاد می‌شوند. ویرایشگرهای پایه توسط یک gRNA به یک مکان خاص قرار می‌گیرند و می‌توانند سیتیدین را در یک پنجره ویرایش کوچک در نزدیکی سایت PAM به یوریدین تبدیل کنند. اوریدین متعاقباً از طریق ترمیم برش پایه به تیمیدین تبدیل می‌شود و تغییر C به T (یا G به A در رشته مخالف) ایجاد می‌کند.

به طور مشابه، ویرایشگرهای پایه آدنوزین برای تبدیل آدنوزین به اینوزین مهندسی شده‌اند، که سلول با آن مانند گوانوزین رفتار می‌کند و تغییر A به G (یا T به C) ایجاد می‌کند. آدنین DNA دامینازها در طبیعت وجود ندارند، اما با تکامل هدایت شده *Escherichia coli* TadA، یک tRNA آدنین دامیناز ایجاد شده‌اند. مانند ویرایشگرهای پایه سیتوزین، دامنه تکامل یافته TadA با پروتئین Cas9 ترکیب می‌شود تا ویرایشگر پایه آدنین ایجاد شود.

هر دو نوع ویرایشگر پایه با چندین نوع Cas9 از جمله Cas9 با ثبات بالا در دسترس هستند. پیشرفت‌های بیشتری با بهینه‌سازی بیان پروتئین، اصلاح ناحیه پیوندی بین نوع Cas و دامیناز برای تنظیم پنجره ویرایش، یا افزودن ترکیب‌هایی که خلوص محصول را افزایش می‌دهند مانند مهارکننده DNA گلیکوزیلاز (UGI) یا Gam مشتق از باکتریوفاژ (Mu-GAM) انجام شده است.

#### 5.4.1 انواع کریسپر

طبقه‌بند rova و همکاران ۵ نوع سیستم کریسپر را تعریف می‌کند که دارای ۱۶ زیر نوع بر اساس ویژگی‌های مشترک و شباهت تکاملی است. اینها به دو دسته بزرگ تقسیم می‌شوند. کلاس‌ها بر اساس ساختار پیچیده‌ای است که DNA ژنوم را تجزیه می‌کند. نوع II CRISPR/Cas اولین سیستم برای مهندسی ژنوم، با نوع V در ۲۰۱۵ بود. [۴]

در گام بعدی از روی ژن‌های کمپلکس cas هم پروتئین Cas9 ساخته می‌شود. سپس کمپلکس Cas9-crRNA-tracrRNA تشکیل می‌شود؛ که این کمپلکس لازم و ضروری برای هدف قرار دادن یا تخریب خارجی DNA می‌باشد.

#### (Nick) Break Single-Strand

در حالی که بسیاری از ویرایشگرهای پایه برای کار در یک پنجره بسیار نزدیک به دنباله PAM طراحی شده‌اند، برخی از سیستم‌های ویرایش پایه طیف گسترده‌ای از انواع تک نوکلئوتیدی (somatic hypermutation) را در یک پنجره ویرایش گسترده‌تر ایجاد می‌کنند و بنابراین برای تکامل هدایت‌شده مناسب هستند. برنامه‌های کاربردی نمونه‌هایی از این سیستم‌های ویرایش پایه عبارتند از جهش‌زایی هدفمند با واسطه AID (TAM) و CRISPR-X، که در آن Cas9 با سیتیدین دامیناز (AID) ناشی از فعال‌سازی ترکیب می‌شود.

نیکاز CRISPR/Cas جهش‌یافته، به جای شکستگی‌های دو رشته‌ای ایجاد شده توسط آنزیم‌های Cas، شکستگی‌های تک‌رشته‌ای با هدف gRNA را در DNA ایجاد می‌کنند. برای استفاده از جهش نیکاز، به دو gRNA نیاز دارید که رشته‌های مخالف DNA شما را در مجاورت یکدیگر مورد هدف قرار دهند. این شیارهای دوتایی یک شکست دو

رشته‌ای (DSB) ایجاد می‌کنند که با استفاده از اتصال انتهایی غیر همولوگ (NHEJ) و مستعد خطا تعمیر می‌شود. استراتژی‌های دوتایی اثرات ناخواسته off-targets را کاهش می‌دهند. جهش‌یافته‌های نیکاز همچنین می‌توانند با یک الگوی تعمیر برای معرفی ویرایش‌های خاص از طریق تعمیر هدایت‌شده همولوژی (HDR) استفاده شوند.

در حالی که Cas9 *S. pyogenes* (SpCas9) مطمئناً متداول‌ترین اندونوکلاز CRISPR برای مهندسی ژنوم است، ممکن است بهترین اندونوکلاز برای هر کاربرد نباشد. به عنوان مثال، توالی PAM برای SpCas9 (5'-NGG-3') در سراسر ژنوم انسان فراوان است، اما یک توالی NGG به درستی برای هدف قرار دادن ژن‌های مورد نظر برای اصلاح قرار نگیرد. این محدودیت در هنگام تلاش برای ویرایش یک ژن با استفاده از تعمیر هدایت‌شده همولوژی (HDR) که نیاز به توالی‌های PAM در مجاورت بسیار نزدیک به منطقه برای ویرایش دارد، نگران‌کننده است.

برای رسیدگی به این محدودیت‌ها، محققان آنزیم‌های SpCas9 را با ویژگی‌های تغییر یافته PAM با استفاده از روش‌های مختلفی از جمله تکامل به کمک فاژ و جهش‌زایی هدایت‌شده مهندسی کرده‌اند. این منجر به توسعه چندین نوع مشتق شده از SpCas9 با توالی‌های PAM غیر NGG شد. جایگزین دیگر Cas9، xCas9 است که مجموعه وسیعی از توالی‌های PAM مانند GAA، NG و GAT را هدف قرار می‌دهد، در حالی که حداقل فعالیت خارج از هدف را نیز نشان می‌دهد.

مراجع این فصل: [2, 3, 30-44, 50]

شکل 8.1: مکانیزم TALEN



اصطلاح	تعریف
ویرایشگر پایه (Base editor)	ادغام یک پروتئین Cas به یک دامیناز که تبدیل مستقیم باز در RNA یا DNA را بدون شکست دو رشته DNA امکان پذیر می کند.
Cas	Cas12a و Cas9 شامل نوکلئازهایی مانند CRISPR Associated Protein, (همچنین به عنوان Cpf1 شناخته می شود)
CRISPR	تناوب‌های کوتاه پالیندروم فاصله‌دار منظم خوشه‌ای، یک منطقه ژنومی باکتریایی که در دفاع از پاتوژن استفاده می شود
CRISPRa	CRISPR Activation; استفاده از فعال کننده dCas9 یا dCas9 با gRNA برای افزایش رونویسی یک ژن هدف
CRISPRi	CRISPR Interference; استفاده از dCas9 یا سرکوبگر-dCas9 با gRNA برای مانع/کاهش رونویسی یک ژن هدف
برش	شکستن دو رشته ای DNA
dCas9	Nuclease dead Cas9, شکل آنزیمی غیر فعال Cas9. می تواند متصل شود، اما نمی تواند DNA را بشکند
جفت نیکاز یا نیک دوتایی (Dual nickase/Double nick)	روشی برای کاهش اثرات خارج از هدف با استفاده از یک نیکاز Cas9 و gRNA 2 مختلف که در مجاورت رشته های مخالف DNA متصل می شوند تا یک DSB ایجاد کنند.
اصلاح یا ویرایش ژنتیکی (Genetic modification or manipulation)	هر گونه اختلال ژنتیکی، از جمله حذف ژنتیکی، فعال سازی ژن، یا سرکوب ژن
gRNA	Guide RNA, باکتریایی درون‌زا که از ادغام مصنوعی crRNA و tracrRNA به وجود می‌آید که هم هدف و هم امکان چسبیدن به Cas9 فراهم می‌کند. این ادغام مصنوعی در طبیعت وجود ندارد و معمولاً به آن sgRNA نیز می‌گویند.
gRNA scaffold sequence	توالی درون gRNA که مسئول اتصال به Cas9 است، شامل توالی هدف/spacer 20 جفت باز که برای هدایت Cas9 به DNA هدف استفاده می‌شود، نمی‌شود.
gRNA targeting sequence	۲۰ نوکلئوتید قبل از توالی PAM در DNA ژنومی قرار دارند. این توالی در یک پلاسمید بیان gRNA کلون می شود اما شامل توالی PAM یا توالی scaffold gRNA نمی شود.
HDR	Homology Directed Repair, یک مکانیسم ترمیم DNA که از یک الگو برای ترمیم نیک های DNA یا DSB ها استفاده می کند
این‌دل (Indel)	Insertion/deletion, نوعی جهش که می تواند منجر به اختلال در یک ژن با جابجایی ORF و/یا ایجاد کدون های توقف زودرس شود.
NHEJ	Non-Homologous End Joining; مکانیزم ترمیم DNA که اغلب باعث می‌شود که این‌دل‌ها به وجود بیایند.
نیک (Nick)	شکست تنها در یک رشته dsDNA
Nickase	Cas9 با یکی از دو حوزه نوکلئاز غیرفعال شده است. این آنزیم قادر است تنها یک رشته از dsDNA هدف را جدا کند.
اثرات off-target یا فعالیت off-target	برش Cas9 در مکان های نامطلوب به دلیل توالی هدف gRNA با همولوژی کافی برای جذب Cas9 در مکان‌های ژنومی ناخواسته
فعالیت On-target	برش Cas9 در محل مورد نظر مشخص شده توسط یک توالی هدف gRNA
ORF	Open Reading Frame; کدون های ترجمه شده که یک ژن را می سازند
PAM	Protospacer Adjacent Motif; توالی مجاور توالی هدف که برای اتصال آنزیم های Cas به DNA هدف ضروری است
PCR	Polymerase Chain Reaction; برای تقویت و خوانا شدن یک توالی خاص از DNA استفاده می شود
مکان هدف	هدف ژنومی gRNA این توالی شامل هدف منحصر به فرد ۲۰ جفت باز مشخص شده توسط gRNA به همراه توالی PAM ژنومی است.

Species/Variant of Cas9	PAM Sequence*
<i>Streptococcus pyogenes</i> (SP); SpCas9	3' NGG
SpCas9 D1135E variant	3' NGG (reduced NAG binding)
SpCas9 VRER variant	3' NGCG
SpCas9 EQR variant	3' NGAG
SpCas9 VQR variant	3' NGAN or NGNG
xCas9	3' NG, GAA, or GAT
SpCas9-NG	3' NG
<i>Staphylococcus aureus</i> (SA); SaCas9	3' NNGRRT or NNGRR(N)
<i>Acidaminococcus</i> sp. (AsCpf1) and <i>Lachnospiraceae</i> bacterium (LbCpf1)	5' TTTV
AsCpf1 RR variant	5' TYCV
LbCpf1 RR variant	5' TYCV
AsCpf1 RVR variant	5' TATV
<i>Campylobacter jejuni</i> (CJ)	3' NNNNRYAC
<i>Neisseria meningitidis</i> (NM)	3' NNNNGATT
<i>Streptococcus thermophilus</i> (ST)	3' NNAGAAW
<i>Treponema denticola</i> (TD)	3' NAAAAC

جدول 1.1: R = G or A, Y = C or T, W = A or T, N = A or C or G or T



## فصل 2

### Literature the of Review

مطالعات زیاد و متعددی روی مشکلات **crispr** انجام شده است ولی در اینجا ما آن‌ها را به دو دسته مختلف تقسیم می‌کنیم، روش‌های مستقیم که در آن‌ها دانشمند به رابطه‌های مستقیم بین مکانیزم‌ها مختلف و تاثیر آنها روی دقت و حساسیت طرح‌ها مورد بررسی قرار دادند. و دسته دوم که روش‌های یادگیری ژرف برای پیش‌بینی تاثیر و حساسیت طرح‌ها.

#### 1.2 روش‌های مستقیم

##### 1.1.2 Chopchop [29-27]

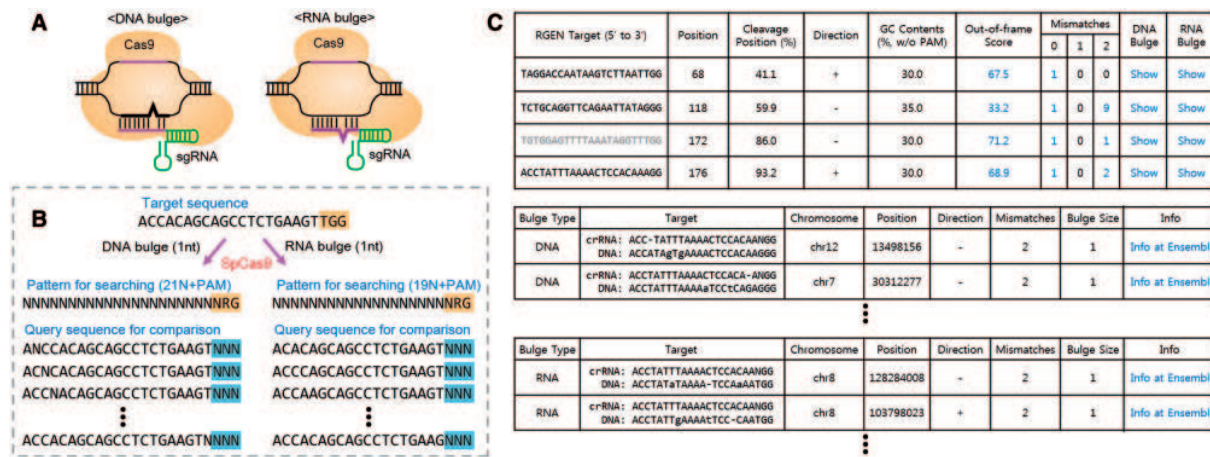
این مقاله که الگوریتم خود را سه بار بروزرسانی کرده است، به عنوان ورودی رشته DNA ورودی و یا اسم ژن یا مختصات آن را می‌گیرد هم چنین مورد استفاده‌ی طرح را می‌پرسد. به عنوان خروجی لیست مرتب شده طرح‌ها ممکن را به هم راه offtargets های آن را به ما پس می‌دهد. برای پیدا کردن offtarget از الگوریتمی به نام bowtie استفاده می‌کنند و primer3 برای پیدا کردن primer ها استفاده می‌کند، این الگوریتم با توجه به پژوهش‌های قبلی از ۶ ویژگی مهم برای مرتب کردن طرح‌ها استفاده می‌کنند که عبارت اند از: تعداد offtarget ها، معماری ژن، GC-Content، وجود نوکلئید G در ۲۰ امین نقطه طرح و همین طور مکان هدف در ژن در ورژن دو این الگوریتم، خروجی روی UCSC هم دیده می‌شود و در مورد PAM استفاده شده در طرح کاربر اختیار بیشتر دارد و میتواند از طرح‌های مختلف CAS استفاده کند. در این ورژن الگوریتم مرتب سازی برحسب حساسیت و تاثیر طرح‌ها است. در زمان بین ورژن یک و دو الگوریتم، دانشمندان به نتایج زیر رسیدند که همه در الگوریتم chopchop موثر هستند: قابل درستی بودن هدف در احتمال شکسته شدن DNA تاثیر مثبت دارد، به همین دلیل این تاثیر را از تاثیر مکان و ترکیب تشکیل دهنده‌ی طرح جدا کردند. میزان خود مکمل بودن طرح در دقت آن تاثیر مستقیم دارد پس برای آن یک امتیاز درست کردن که خب بر حسب مکمل بودن دو دویی نوکلئوتیدها اول آخر طرح است. و در انتها این امتیازهای جدید را با SVM و متریک‌ها مختلف برای مرتب سازی طرح استفاده کردند و اسم آن را امتیاز تاثیر قرار دادند. برای تعیین حساسیت هر طرح الگوریتم از دست‌آوردهای جدید پژوهشگرها استفاده کردند: استفاده از دو طرح برای شکستن یک رشته DNA، عدم تطابق در PAM هم به عنوان offtarget محسوب میشود و حتی در بعضی طرح‌ها باعث حساسیت بهتر می‌شود، یک عدم تطابق در ۱۱bp از سمت ۵' و یا داشتن بیشتر از ۴ عدم تطابق باعث شکسته نشدن DNA و کوتاه کردن طول sgRNA باعث حساسیت بهتر می‌شود، با توجه به این اطلاعات offtarget ها با bowtie2 پیدا می‌کند و با توجه به آنها امتیاز حساسیت می‌دهد.

##### 2.1.2 Cas-Designer و Cas-OFFinder [26,25]

این دو الگوریتم به دنبال پیدا کردن بهترین sgRNA و مناطق off-target یک ژنوم مشخص یا توالی‌های تعریف شده توسط کاربر هستند. Cas-Designer، یک برنامه کاربرپسند برای کمک به محققان در انتخاب مناسب مکان‌های هدف در یک ژن مورد علاقه برای RNA مشتق شده از CRISPR/Cas نوع II، که در حال حاضر به طور گسترده برای تحقیقات زیست پزشکی و بیوتکنولوژی استفاده می‌شود. Cas-Designer به سرعت ارائه می‌دهد فهرستی از تمام توالی‌های RNA راهنمای ممکن در یک توالی DNA ورودی داده شده و آنها off-target در ژنوم انتخابی. علاوه بر این، برنامه امتیاز خارج از چارچوب را به هر سایت هدف اختصاص می‌دهد تا به کاربران کمک کند سایت‌های مناسب برای ژن را برای Knockout انتخاب کنند. Cas-Designer نتایج را در یک جدول تعاملی نشان می‌دهد و فیلتر کاربر پسند را ارائه می‌دهد کارکرد.

ابتدا Cas-Designer سایت‌های طرح‌های احتمالی را با یک کاربر تعریف شده [50-NGG-30 یا 50-NRG-30 برای SpCas9، 50-NNAGAAW-30 برای StCas9 (Cong et al., 2013)، 50-NNNGMTT-30 برای NmCas9 (Hou et al., 2013) و 50-NNGRRT-30 برای SaCas9 (Ran et al., 2015)] در یک توالی DNA معین پیدا می‌کند. دوم، Cas-Designer امتیاز خارج از قاب مرتبط با





شکل 1.2: (الف) شماتیک مکان‌های off-targets را با برآمدگی DNA یا RNA نشان می‌دهد. (ب) استراتژی برای برآمدگی 1-nt DNA یا RNA بر اساس Cas-OffFinder (ج) یک مثال از یک جدول خروجی Cas-Designer تمام gRNA های ممکن را از توالی های ورودی به همراه اطلاعات مفید (بالا) نشان می‌دهد. اگر کاربر روی رنگ آبی کلیک کند عدد، کلمه یا عبارت، اطلاعات دقیق تری مانند اهداف برآمدگی DNA (وسط) یا RNA (پایین) ارائه می‌شود. علاوه بر این، کاربر می‌تواند موارد مربوطه را به دست آورد اطلاعات ژنومی از طریق مرورگر ژنوم Ensembl (Flicek و همکاران، 2011)، با کلیک بر روی دکمه "اطلاعات در Ensembl"

میکروهمولوژی را به سرعت محاسبه می‌کند که با فراوانی جهش‌های تغییر قاب همبستگی مثبت دارد (Bae et al., 2014b). محتوای GC و امتیازات خارج از کادر در این مرحله موقعیت‌های برش را نشان می‌دهد.

Cas-OffFinder از دو هسته OpenCL مختلف تشکیل شده است هسته جست‌وجوگر و یک هسته مقایسه‌گر) و C++ ابتدا Cas-OffFinder فایل‌های داده توالی ژنوم را به صورت تک یا چندتایی در فرمت FASTA می‌خواند. سپس در هسته جست‌وجو بارگذاری می‌شود که تمام سایت‌هایی را که شامل یک توالی PAM در کل ژنوم هستند، کامپایل می‌کند. برای جست‌وجو و انتخاب سریع و مؤثر این سایت‌های خاص، هسته جست‌وجوگر به‌طور مستقل روی هر واحد محاسباتی یک پردازنده اجرا می‌شود، یعنی همه فرآیندهای جست‌وجو در واحدهای محاسباتی به‌طور همزمان انجام می‌شوند.

### 3.1.2 E-CRISP [24]

در اینجا ما E-CRISP، یک برنامه وب برای طراحی توالی‌های gRNA را توصیف می‌کنیم. (الف) مراحل E-CRISP. ابتدا کاربر ارگانیزم و دنباله هدف را انتخاب می‌کند. این هدف می‌تواند یک نماد ژن، یک شناسه ENSEMBL یا یک توالی FASTA باشد. دوم، کاربر هدف آزمایش ویرایش را مشخص می‌کند. بسته به هدف، E-CRISP مناطق مختلفی از توالی ژن را مورد هدف قرار می‌دهد. سوم، E-CRISP نتایج را با توجه به اطلاعات حاشیه نویسی ژن فیلتر می‌کند. چهارم، اهداف خارج از هدف بر اساس تراز توالی هر طرح با ژنوم مرجع تجزیه و تحلیل می‌شوند. در نهایت، E-CRISP یک صفحه خروجی تعریف شده توسط کاربر تولید می‌کند. (ب) RNA های راهنما در برابر جایگاه let-7 گونه‌های مشخص شده طراحی شده‌اند. توالی و محل gRNA های بالغ از miRBase بازبایی شده است. این خروجی انعطاف‌پذیر و پارامترهای طراحی آزمایش‌گرا را فراهم می‌کند، طراحی کتابخانه‌های متعدد و در نتیجه تجزیه و تحلیل سیستماتیک تأثیر پارامترهای مختلف را ممکن می‌سازد. E-

شکل 2.2: الگوریتم E-CRISP

CRISP توالی‌های هدف مکمل gRNA را شناسایی می‌کند که به یک موتیف که از سمت 3' مجاور به (A یا G) N ختم می‌شود، که برای هسته Cas9 مورد نیاز است تا رشته دوگانه DNA را برش دهد. E-CRISP از یک رویکرد نمایه سازی سریع برای یافتن مکان های اتصال و یک درخت فاصله دودویی برای حاشیه نویسی سریع سایت های هدف gRNA احتمالی استفاده می‌کند. با استفاده از این الگوریتم‌ها، می‌توان در چند ساعت کتابخانه‌هایی در مقیاس ژنومی برای چندین موجود زنده ایجاد کرد.

## 4.1.2 CRISPOR [23]

## Guides transcribed in cells from a U6 promoter

Wang/Xu HL60 (2076)	0.616	0.343	0.486	0.321	0.246	0.201	0.485
Doench 2014 Mouse-EL4 (951)	0.427	0.577	0.400	0.403	0.369	0.156	0.700
Koike-Yusa/Xu 1 M-ESC (907)	0.281	0.221	0.306	0.12	0.119	0.094	0.367
Chari 293T (1234)	0.310	0.246	0.286	0.457	0.308	0.123	0.381
Doench 2016 A375 (2333)	0.265	0.266	0.287	0.245	0.164	0.144	0.540
Hart Repl2Lib1 Hct116 (4239)	0.307	0.288	0.292	0.208	0.232	0.159	0.384
Gandhi Electrop. Ciona (72)	0.298	0.245	0.150	0.248	0.112	0.354	0.419
Farboud <i>C. elegans</i> (50)	0.476	0.301	0.545	0.602	0.400	0.177	0.541
Ren <i>Drosophila</i> (39)	0.313	0.178	0.225	0.152	-0.158	-0.347	0.131

Guides transcribed *in vitro* from a T7 promoter

Varshney Zebrafish (102)	0.17	0.139	0.171	0.28	0.27	0.262	0.219
Gagnon Zebrafish (111)	0.207	-0.072	0.179	0.202	0.083	0.357	0.104
Moreno-Mateos Z-fish (1020)	0.14	0.038	0.171	0.145	0.037	0.579	0.12

Color Key  
-0.5 0.5  
Value

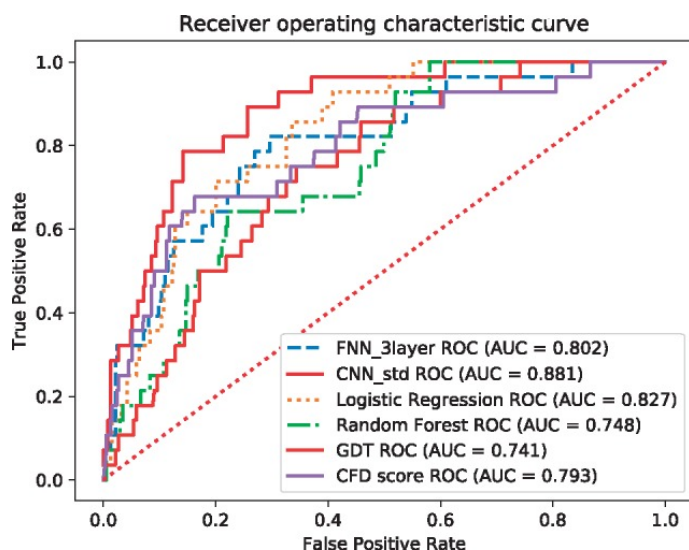
Wang Score  
Doench Score  
Xu Score  
Chari Score  
Wang Score  
Moreno-Mateos Score  
Fuell/Doench Score

شکل 3.2: هم بستگی اسپیرمن بین امتیاز موثر بودن و داده‌ها

## 2.2 روش‌های یادگیری ژرف

## 1.2.2 پیش‌بینی off-target به کمک یادگیری ژرف

پیش‌بینی جهش‌های خارج از هدف در CRISPR-Cas9 به دلیل ارتباط آن با تحقیقات ویرایش ژن یک موضوع پرپژوهشی است. روش‌های پیش‌بینی مختلفی توسعه یافته‌اند. با این حال، اکثر آنها فقط امتیازات را بر اساس عدم تطابق با دنباله راهنما در CRISPR-Cas9 محاسبه کردند. بنابراین، روش‌های پیش‌بینی موجود قادر به مقیاس‌بندی و بهبود عملکرد خود با گسترش سریع داده‌های تجربی در CRISPR-Cas9 نیستند. علاوه بر این، روش‌های موجود هنوز نمی‌توانند دقت کافی را در پیش‌بینی‌های خارج از هدف برای ویرایش ژن در سطح بالینی برآورده کنند. برای رفع این مشکل، مقاله دو الگوریتم را با استفاده از شبکه‌های عصبی عمیق برای پیش‌بینی جهش‌های off-target در ویرایش ژن CRISPR-Cas9 طراحی و پیاده‌سازی می‌کنیم (به توجه به اطلاعات اولین الگوریتم ماشینی). این مدل‌ها بر روی مجموعه داده‌های اخیراً منتشر شده، مجموعه داده‌های CRISPOR، برای معیار عملکرد، آموزش دیده و آزمایش شدند. یکی دیگر از مجموعه داده شناسایی شده توسط GUIDE-seq برای ارزیابی بیشتر مورد استفاده قرار گرفت. مقاله نشان می‌دهد که شبکه عصبی کانولوشن بهترین عملکرد را در مجموعه داده‌های CRISPOR به دست می‌آورد، و سطح طبقه‌بندی متوسط زیر منحنی ۰.۹۷ درصد را تحت اعتبارسنجی متقاطع 5 برابری طبقه‌بندی شده به دست می‌آورد. جالب اینجاست



شکل 4.2: هم بستگی اسپیرمن بین امتیاز موثر بودن و داده‌ها [21]

که شبکه عصبی پیشخور عمیق نیز می‌تواند با میانگین ۰.۹۷ در همان تنظیمات رقابتی باشد. ما دو مدل شبکه عصبی عمیق را با روش‌های پیشرفته پیش‌بینی off-target (مانند CFDP، MIT، CROP-IT و CCTop) و سه مدل سنتی یادگیری ماشینی (یعنی جنگل تصادفی، درخت‌های تقویت‌کننده گرادین، و رگرسیون لجستیک) در هر دو مجموعه داده از نظر مقادیر AUC نشان دهنده لبه‌های

رقابتی الگوریتم‌های پیشنهادی است. تحلیل‌های اضافی برای بررسی دلایل زمینه‌ای از دیدگاه‌های مختلف انجام می‌شود.

### 2.2.2 CCTop [19]

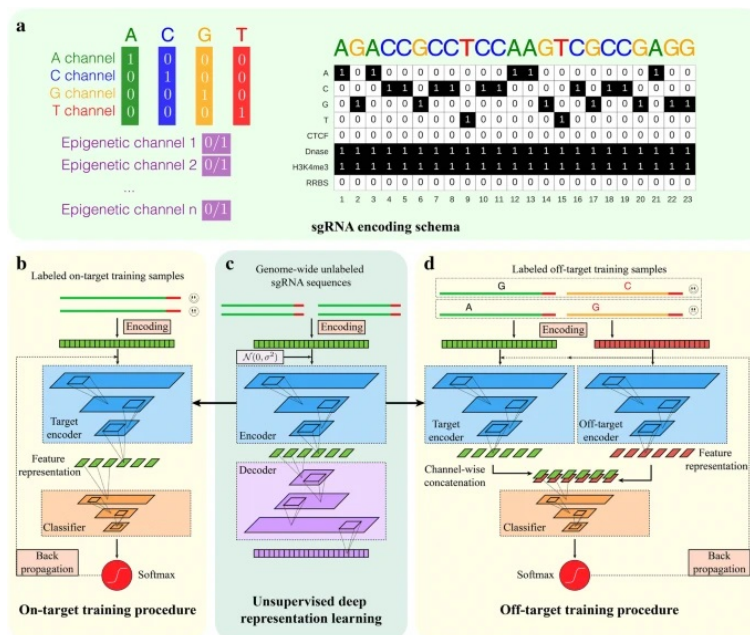
این روش برای اینکه طرح‌های مختلف که به صورت N20NGG هستند را دسته‌بندی کند، ابتدا با آزمایش‌های عملی طرح‌ها را به دو کلاس موثر و ناموثر دسته‌بندی کرد. آزمایش به این گونه بود که در محیط آزمایشگاهی طرح را به ژن طزیق می‌کردند و ادامه آن پیشتر هر طرح را با شمردن هدف‌های تغییر کرده در طول زمان را یاد داشت کرده. این روش بر این باور بود که ribosomal و non-ribosomal بودن ژن در تاثیر طرح موثر است پس دیتاست خود را به دو قسمت تقسیم کرده و برای طرح هر کدام sgRNA موثر و ناموثر را تایین کرده. این طرح جایگاه هر نیکلوتاید را در های sgRNA موثر و ناموثر بررسی کرده و به نتایج زیر رسیده است. نحوی انتخاب موثر یا ناموثر بودن یک طرح با کمک مدل حسب مدل Elastic-Net است که در آن اگر  $X_i$  encode شده طرح‌ها باشند و  $Y$  امتیاز آن‌ها باشد داریم:

پس از تمرین این مدل به ۲۸ ویژگی رسید که بیشتر ویژگی‌ها در ناحیه اسپیسر واقع شده اند و بعضی از آنها قبلاً پیدا شده بود و بعضی جدید بود:

- ها G به شدت در موقعیت‌های 1- و 2- به PAM در CAS9 ترجیح داده می‌شوند
- ها T در چهار موقعیت نزدیک به PAM نامطلوب هستند
- نوکلئوتیدهای از ۵' به ۳'، در حالی که توالی در سمت برعکس تاثیر قابل توجهی ندارد.
- ها C در موقعیت 3- در CAS9 ترجیح داده می‌شوند
- ها A در موقعیت 5- تا 12- ترجیح داده می‌شود.
- ها G در موقعیت‌های 14- تا 17- ترجیح داده می‌شوند.

### 3.2 DeepCRISPR [22]

DeepCRISPR یک پلت فرم محاسباتی جامع برای یکپارچه سازی پیش‌بینی سایت sgRNA روی هدف و خارج از هدف در یک چارچوب با یادگیری عمیق، پیشی گرفتن از پیشرفته‌ترین ابزارهای موجود در سیلیکون. DeepCrispr [37] علاوه بر ویژگی‌های توالی DNA، چهار ویژگی اپی ژنتیکی را معرفی کرد و به طور خودکار اطلاعات معتبر را با استفاده از اصل Auto-encoder استخراج می‌کند. چندین مدل از جمله برش هدف sgRNA و پیش‌بینی تمایل خارج از هدف ایجاد شد. این پژوهشگران بر این باور بودند که خود دنباله sgRNA می‌تواند اطلاعات مفید درباره موثر بودن یک توالی sgRNA بدهد به همین امر مدل خود را به دو گونه آموزش دادن با استفاده از اطلاعات اپی ژنتیکی و با اطلاعات اپی ژنتیکی که نشان می‌دهد که اطلاعات اپی ژنتیکی بی‌تاثیر نیست.



شکل 5.2: DeepCRISPR





## فصل 3

### Methods



#### 1.3 ensemble

در آمار و یادگیری ماشین، روش‌های ensemble از الگوریتم‌های یادگیری چندگانه استفاده می‌کنند تا عملکرد پیش‌بینی‌کننده بهتری نسبت به هر یک از الگوریتم‌های یادگیری سازنده به‌تنهایی به‌دست آورند. [7-9] بر خلاف ensemble آماری، که معمولاً از بی‌نهایت مکانیک آماری استفاده میکند، یک مجموعه یادگیری ماشینی تنها از مجموعه محدود مشخصی از مدل‌های تشکیل شده است، اما معمولاً ساختار بسیار انعطاف‌پذیرتری را در بین آن گزینه‌ها امکان می‌دهد.

##### 1.1.3 تعریف

الگوریتم‌های یادگیری نظارت شده وظیفه جستجو در فضای فرضیه را برای یافتن یک فرضیه مناسب انجام می‌دهند که پیش‌بینی‌های خوبی را با یک مسئله خاص انجام دهد. [10]

ارزیابی پیش‌بینی یک مجموعه معمولاً به محاسبات بیشتری نسبت به ارزیابی پیش‌بینی یک مدل نیاز دارد. از یک جهت، یادگیری گروهی ممکن است به عنوان راهی برای جبران الگوریتم‌های یادگیری ضعیف با انجام محاسبات زیاد در نظر گرفته شود. از سوی دیگر، جایگزین این است که یادگیری بسیار بیشتری را در یک سیستم غیر گروهی انجام دهید. یک سیستم ensemble ممکن است در بهبود دقت کلی برای افزایش یکسان در منابع محاسباتی، ذخیره‌سازی یا ارتباطی با استفاده از این افزایش در دو یا چند روش، کارآمدتر از افزایش استفاده از منابع برای یک روش واحد باشد. الگوریتم‌های سریع مانند درخت‌های تصمیم معمولاً در روش‌های ensemble (مثلاً جنگل‌های تصادفی) استفاده می‌شوند، اگرچه الگوریتم‌های کندتر می‌توانند از تکنیک‌های مجموعه نیز بهره ببرند.

برای اینکه بتوان از این روش استفاده کرد نیاز است که ابتدا جواب این مدل‌ها یا اکسپرت‌ها را روی یک دیتای مشابه داشته باشیم، مقاله‌ی DeepCRISPR دقیقاً داده ۴۲۵ دنباله sgRNA از امتیاز دهنده‌های ۵ مقاله و امتیاز مقاله خود تهیه کرده که از آنها استفاده کردیم. چندین روش ensemble برای جمع این امتیازها و رتبه‌بندی‌ها استفاده کردیم، مانند وزن دهی بر حسب دقت هر مدل روی یک دیتا ثابت و همین‌طور روش LPA یا Latent Profile Analysis که به ما مدلی برحسب پیش‌بینی مدل‌هایی دیگر می‌دهند. از این روش‌ها ما دو مدل بدست آوردیم ولی دقت این مدل‌ها همگی از مدل DeepCRISPR پایین‌تر بودند با آنالیز بیشتر به این نتیجه رسیدیم که این مدل‌ها بر سر بعضی نقاط شدیداً اختلاف نظر دارند که باعث تاثیر منفی در نتیجه ensemble این مدل‌ها می‌شود و با این گونه وزن دهی نمی‌توان به نتیجه بهتری رسید.

در مرحله بعدی با جنگل‌های تصادفی سعی کردیم کردیم فضای مسئله را تقسیم کنیم و بر اساس آن از امتیاز مدل‌های دیگر استفاده کنیم تا بتوانیم جواب بهتری بدست آوریم، پس از تنظیم کردن ابرپارامترها به توانستیم به مدلی بهتر از مدل‌های قبلی برسیم ولی با انجام cross-validation به این نتیجه رسیدیم که دیتای استفاده شده برای آموزش تاثیر زیادی روی دقت پیش‌بینی دارد و لزوماً این روش همیشه از روش Deepcrispr بهتر نیست، برای بدست آوردن مدل قوی نیاز به دیتای بیشتر داشتیم.

در مرحله‌ی آخر، با توجه به اینکه اکسپرت‌ها اختلاف نظر داشتند و ensemble کردن این اکسپرت‌ها اختلاف نظر آنها را کم می‌کرد، چهار الگوریتم، RandomForestRegressor، LinearRegression، GradientBoostingRegressor، ExtraTreeRegressor را برای ensemble اکسپرت‌ها انتخاب کردیم. هر کدام از الگوریتم‌ها به تنهایی به داده آموزش حساس بودند و با انجام cross-validation لزوماً به نتیجه بهتری نمی‌رسیدند ولی برخلاف اکسپرت‌های اولیه اختلاف نظر این رگرسورها خیلی کم بود و پس این متدها را با هم ادغام و به نتیجه‌ی مطلوب رسیدیم، یعنی مدلی به دیتا حساس نبود و با هر فولدی باز هم از روش DeepCRISPR بهتر عمل می‌کرد.

برای اینکه مشکل داده کم را حل کنیم، ما ابتدا ۲۷۰۵ توالی مختلف را در الگوریتم‌های CRISPOR E-Crisp، CCTop، Cas-OFFinder،

و Chopchop جمع آوری کردیم، که منجر بدست آمدن ۵۰ هزار sgRNA یکتا شد و از آنجا که خروجی الگوریتم‌ها می‌توانست NaN هم باشد، با حذف این داده‌ها به ۳۶ هزار sgRNA یکتا و نظر اکسپرت‌ها راجع به آن رسیدیم. تنها کافی بود که بتوانیم یک golden standard برای این داده‌ها پیدا کنیم، که متاستفانه قادر به این کار نشدیم.

## 2.3 Attention

موفقیت ما در روش ensemble، برخلاف الگوریتم‌های دیگر که با استفاده از اطلاعات جانبی دیگر در مورد sgRNA بود، بر حسب نمایش دادن sgRNA در یک بردار معنا دار بود. به همین جهت ما سعی کردیم که یک بردار معنا دار از هر sgRNA بسازیم و برای این امر از روش توجه استفاده کردیم.

در شبکه‌های عصبی، توجه تکنیکی است که توجه شناختی را تقلید می‌کند. این اثر باعث می‌شود که اثر برخی از بخش‌های ورودی افزایش یابد در حالی که بخش‌های دیگر را کاهش می‌دهد - فکر این است که شبکه باید تمرکز بیشتری را به آن بخش کوچک اما مهم داده اختصاص دهد. یادگیری اینکه کدام بخش از داده‌ها مهم‌تر از سایرین است بستگی به زمینه دارد و با نزول گرادین آموزش داده می‌شود.

مکانیسم‌های مانند توجه در دهه ۱۹۹۰ با نام‌هایی مانند ماژول‌های ضربی، واحدهای سیگما پی و ابرشبکه‌ها معرفی شدند. [11] انعطاف‌پذیری آن ناشی از نقش آن به عنوان "وزن نرم" است که می‌تواند در طول زمان اجرا تغییر کند، برخلاف وزنه‌های استاندارد که باید در زمان اجرا ثابت بمانند. کاربردهای توجه شامل حافظه در ماشین‌های تورینگ عصبی، وظایف استدلال در رایانه‌های عصبی متمایز [12]، پردازش زبان در ترانسفورماتورها، و پردازش داده‌های چندحسی (صدا، تصاویر، ویدئو، متن) در درک‌کننده‌ها است. [13-17]

این مدل‌ها از دو قسمت نظارت شده و نظارت نشده تشکیل شده که اولین آموزش برای پیدا کردن ساختار کلی است و دومین آموزش برای تنظیم مناسب برای امر خاص است.

در اینجا ما چند مدل مختلف مانند bert و roberta و DNAbert برای کلاس بندی sgRNA استفاده کردیم که نتایج این مدل‌ها خیلی ضعیف بود. با توجه به آنالیزهای انجام شده به این نتیجه رسیدیم که مشکل از دیتاهای بدون لیب و با لیب استفاده شده در طول آموزش‌ها بود. برای ساخت token ابتدا از روش مرسوم kmer در DNA استفاده کردیم [7] که به این صورت است که برای هر حرف از توالی k حرف بعد از آن تکرار می‌شود. سپس این کلمات تایی k را به عنوان دیکشنری کلمات در نظر می‌گیریم. برای قسمت pretrain از sgRNA که خودمان ذخیره کرده بودیم و داده‌های دیگر استفاده کردیم و سپس برای تنظیمات نهایی از داده‌های مقاله DNAbert استفاده کردیم ولی نتایج آن نتایج جالبی نبود.



## فصل 4

### Results



#### Ensemble 1.0.4

نمونه‌ای از نتایج استفاده مستقیم روش‌های LPA و رگرسیون برای پیدا کردن وزن خوب بین اکسپرت‌ها

AUC_ROC								
	Model 1 Score for Class 1	Model 1 Score for Class 2	reg	DeepCRISPR	CRISPRater	SSC	sgRNA_Scorer	sgRNA_Designer
0	0.298699	0.701301	0.972195	0.937398	0.644634	0.618780	0.651220	0.597073
1	0.283887	0.716113	0.945652	0.927749	0.687002	0.615767	0.639424	0.623890
2	0.262941	0.737059	0.955300	0.927717	0.710353	0.646113	0.659053	0.642117
3	0.259759	0.740241	0.965996	0.903380	0.705412	0.677425	0.644708	0.647344
4	0.275471	0.724529	0.941313	0.831989	0.725124	0.644951	0.649370	0.630279
5	0.308077	0.691923	0.920335	0.756421	0.716509	0.628349	0.618559	0.612786
6	0.351860	0.648140	0.914763	0.739884	0.660604	0.586359	0.580693	0.598542
7	0.368440	0.631560	0.880356	0.684054	0.642937	0.563159	0.568745	0.623188
8	0.478076	0.521924	0.830463	0.635765	0.608747	0.427671	0.475937	0.585022

شکل 1.4: ROC AUC

AUC_PR								
	Model 1 Score for Class 1	Model 1 Score for Class 2	reg	DeepCRISPR	CRISPRater	SSC	sgRNA_Scorer	sgRNA_Designer
0	0.949309	0.984511	0.998971	0.997488	0.980977	0.975903	0.981214	0.979200
1	0.882104	0.960966	0.994282	0.993260	0.958888	0.937486	0.949994	0.945258
2	0.817188	0.944064	0.992143	0.989347	0.941201	0.913606	0.927859	0.921713
3	0.757566	0.922662	0.991772	0.975436	0.919262	0.896279	0.896045	0.897106
4	0.694093	0.888147	0.978102	0.930503	0.890534	0.844805	0.848857	0.836661
5	0.614649	0.817835	0.953671	0.867575	0.834472	0.772920	0.766712	0.768627
6	0.538775	0.715779	0.936461	0.807515	0.714867	0.648901	0.662849	0.673465
7	0.352377	0.531786	0.823824	0.581910	0.520233	0.440453	0.476842	0.494809
8	0.172493	0.185667	0.503301	0.168832	0.148355	0.091240	0.107466	0.179785

شکل 2.4: PR AUC

F1								
	Model 1 Score for Class 1	Model 1 Score for Class 2	reg	DeepCRISPR	CRISPRater	SSC	sgRNA_Scorer	sgRNA_Designer
0	0.785612	0.725309	0.983092	0.980676	0.982036	0.689873	0.980815	0.982036
1	0.734139	0.691928	0.966208	0.925575	0.958333	0.601054	0.957055	0.958333
2	0.683544	0.680484	0.964798	0.832298	0.934837	0.522244	0.933501	0.923077
3	0.645902	0.655678	0.963165	0.694698	0.910256	0.441113	0.911425	0.845188
4	0.575916	0.645914	0.953079	0.399050	0.881720	0.317848	0.876821	0.752108
5	0.518797	0.602564	0.918301	0.113924	0.825545	0.257703	0.825485	0.444444
6	0.458333	0.557377	0.875740	0.000000	0.541463	0.159468	0.762044	0.078571
7	0.366492	0.489297	0.712329	0.000000	0.167488	0.108911	0.579564	0.000000
8	0.161702	0.171123	0.156863	0.000000	0.000000	0.000000	0.200000	0.000000

شکل 3.4: score F1

نمونه‌ای از داده‌های مقاله DeepCRISPR و امتیاز اسپیرمن بین جواب DeepCRISPR و رگرسیون درخت تصادفی

random\_forest: SpearmanResult(correlation=0.5357974645093228, pvalue=7.58034824820603e-12)  
 DeepCRISPR: SpearmanResult(correlation=0.5336854772473561, pvalue=9.56026407893448e-12)

sgRNA_number	KO_reporter_assay	DeepCRISPR_score	CRISPRater_score	SSC_Score	sgRNA_Scorer_score	sgRNA_Designer_rsl_score	sgRNA_sequence	extended_spacer	reg
0	sg1	0.000	0.177065	0.5710	-0.485	30.66	0.571 GAGTCGGGGTTTCGTCATGTTGG	AGTAGAGTCGGGGTTTCGTCATGTTGTCA	0.397722
1	sg2	0.000	0.055157	0.6998	-0.266	54.96	0.533 CGCCGCCGCTTCGGTGATGAGG	CTGCCGCCGCCGCTTCGGTGATGAGGAAA	0.088200
2	sg3	0.000	0.239546	0.6865	-0.448	25.79	0.410 GGCAGCGTCGTGCACGGGTCGGG	CCCGGGCAGCGTCGTGCACGGGTCGGGTGA	0.269884
3	sg4	0.000	0.147778	0.6405	-0.046	53.81	0.491 TGGCGGATCACTTGACGTCAAG	GAGGTGGCGGATCACTTGACGTCAAGAGT	0.175252
4	sg5	0.000	0.120955	0.6796	0.067	12.44	0.485 TTACCATAGTGACGGGTGCAGG	CTTTTACCATAGTGACGGGTGCAGGCAT	0.039664
...	...	...	...	...	...	...	...	...	...
420	sg426	0.953	0.545577	0.7671	0.879	69.61	0.670 GC GTTGGGTGTA CTTCAGAGG	TTGAGCGTTGGGTGTA CTTCAGAGGTGG	0.771596
421	sg427	0.955	0.493218	0.6749	-0.154	13.28	0.555 ATGTAGGGCTTGCGAGTCTAGG	GGATATGTAGGGCTTGCGAGTCTAGGTCA	0.864814
422	sg428	0.955	0.568641	0.7716	0.743	93.33	0.604 GGTAGAAGGTGTA CTCCGGTGG	TCGTGTAGAAGGTGTA CTCCGGTGGGAG	0.913264
423	sg429	0.963	0.173204	0.6069	-0.025	60.36	0.609 GTTACGCAAGTATCATGCATGG	AACAGTTTACGCAAGTATCATGCATGTTTC	0.816500
424	sg430	0.973	0.409570	0.7093	0.801	92.17	0.732 GC GCGGTAGTGTGACCCGGCGG	AGTGGCGGTAGTGTGACCCGGCGGTCC	0.851474

425 rows × 10 columns

شکل 4.4: هم بستگی اسپیرمن بین امتیاز موثر بودن و داده‌ها

نمونه از امتیاز ادغام روش‌های ادغام و داده‌های که برای تست این الگوریتم استفاده شدند.

OURS		State_of_the_Art							
spearman_score:	0.40389303464697907	0.3900903015648729							
MSE_score:	0.05149765311592955	0.10509945128312974							
thershold: 0.7									
accuracy_score:	0.5937500000000000	0.5078125000000000							
roc_auc_score:	0.63520678685047716	0.59199363732767762							
precision_score:	0.8000000000000004	0.82758620689655171							
recall_score:	0.48780487804878048	0.29268292682926828							
f1_score:	0.6060606060606060	0.4324324324324240							
thershold: 0.8									
accuracy_score:	0.5937500000000000	0.6171875000000000							
roc_auc_score:	0.58852258852258854	0.61135531135531140							
precision_score:	0.76190476190476186	0.9375000000000000							
recall_score:	0.25396825396825395	0.23809523809523808							
f1_score:	0.38095238095238093	0.37974683544303794							
thershold: 0.9									
accuracy_score:	0.7656250000000000	0.7500000000000000							
roc_auc_score:	0.53469387755102038	0.48979591836734693							
precision_score:	0.5000000000000000	0.0000000000000000							
recall_score:	0.10000000000000001	0.0000000000000000							
f1_score:	0.16666666666666669	0.0000000000000000							
DeepCRISPR_score	CRISPRater_score	SSC_Score	sgRNA_Scorer_score	sgRNA_Designer_rsl_score	RandomForestRegressor	LinearRegression	*OURS	Golden Standard	
0	0.274436	0.719872	0.833631	0.967399	0.826165	0.663134	0.659704	0.643490	0.306269
1	0.176191	0.139465	0.315245	0.513592	0.637993	0.657013	0.306695	0.348535	0.199383
2	0.550364	0.465483	0.480543	0.261611	0.514337	0.777532	0.648132	0.648823	0.523124
3	0.194020	0.721668	0.375580	0.284382	0.367384	0.662985	0.569410	0.593126	0.862282
4	0.387522	0.490623	0.515530	0.042131	0.654122	0.658408	0.587609	0.623620	0.865365
...	...	...	...	...	...	...	...	...	...
123	0.843723	0.648244	0.694752	0.934397	0.670251	0.784978	0.903759	0.886889	0.927030
124	0.570856	0.358540	0.392360	0.704885	0.517921	0.778118	0.606164	0.609915	0.842754
125	0.340520	0.396449	0.431989	0.572976	0.580645	0.657013	0.511038	0.547437	0.917780
126	0.680362	0.486233	0.555516	0.555823	0.627240	0.779117	0.735950	0.731708	0.770812
127	0.520986	0.542698	0.558729	0.397131	0.584229	0.781099	0.676899	0.663032	0.914697

128 rows × 9 columns

## شکل 5.4: نتیجه آموزش ل

	OURS		State_of_the_Art						
spearman_score:	0.48746235613538696		0.437077149169115						
MSE_score:	0.03988096265673611		0.09215726386787026						
thershold: 0.7									
accuracy_score:	0.6093750000000000		0.4687500000000000						
roc_auc_score:	0.64285714285714279		0.57359307359307365						
precision_score:	0.80357142857142860		0.83333333333333337						
recall_score:	0.53571428571428570		0.23809523809523808						
f1_score:	0.64285714285714279		0.37037037037037035						
thershold: 0.8									
accuracy_score:	0.5781250000000000		0.5546875000000000						
roc_auc_score:	0.56744868035190610		0.54227761485826009						
precision_score:	0.69999999999999996		0.69230769230769229						
recall_score:	0.22580645161290322		0.14516129032258066						
f1_score:	0.34146341463414631		0.24000000000000002						
thershold: 0.9									
accuracy_score:	0.7734375000000000		0.7500000000000000						
roc_auc_score:	0.51395173453996978		0.47058823529411764						
precision_score:	0.28571428571428570		0.0000000000000000						
recall_score:	0.07692307692307693		0.0000000000000000						
f1_score:	0.12121212121212123		0.0000000000000000						
DeepCRISPR_score	CRISPRater_score	SSC_Score	sgRNA_Scorer_score	sgRNA_Designer_rsl_score	RandomForestRegressor	LinearRegression	*OURS	Golden Standard	
0	0.242076	0.457502	0.483042	0.459625	0.573477	0.664506	0.500409	0.535289	0.160329
1	0.420698	0.285914	0.258836	0.195606	0.141577	0.681046	0.470018	0.562233	0.770812
2	0.919773	0.714685	0.753302	0.971612	0.716846	0.782563	0.980870	0.960777	0.856115
3	0.232292	0.458300	0.262406	0.283780	0.195341	0.663113	0.442136	0.504196	0.728674
4	0.675809	0.335595	0.536594	0.617715	0.675627	0.780262	0.694447	0.682980	0.914697
...	...	...	...	...	...	...	...	...	...
123	0.377594	0.388468	0.418065	0.542181	0.648746	0.665039	0.548669	0.607235	0.208633
124	0.297177	0.317638	0.515173	0.746013	0.614695	0.664506	0.483284	0.519896	0.904419
125	0.652755	0.552873	0.324527	0.087672	0.326165	0.780995	0.713469	0.727742	0.817061
126	0.380146	0.536712	0.797929	0.977330	0.838710	0.667273	0.649623	0.656613	0.637205
127	0.616246	0.598763	0.543734	0.541278	0.607527	0.781092	0.754977	0.722380	0.921891

128 rows × 9 columns

## شکل 6.4: هم بستگی اسپیرمن بین امتیاز موثر بودن و داده‌ها



## Attention 2.0.4

نتیجه‌ی دسته بندی بعد از آموزش به کمک مدل‌های توجه

```
06/26/2021 00:48:31 - INFO - __main__ - ***** Eval results *****
06/26/2021 00:48:31 - INFO - __main__ - acc = 0.7098795180722891
06/26/2021 00:48:31 - INFO - __main__ - auc = 0.5
06/26/2021 00:48:31 - INFO - __main__ - f1 = 0.41516347237880497
06/26/2021 00:48:31 - INFO - __main__ - mcc = 0.0
06/26/2021 00:48:31 - INFO - __main__ - precision = 0.35493975903614455
06/26/2021 00:48:31 - INFO - __main__ - recall = 0.5
```

شکل 7.4: نتیجه تمرین به کمک 3mer، به کمک مدل DNAbert

```
mcc = cov_ytyp / np.sqrt(cov_ytyp * cov_ytyp)
06/26/2021 18:49:30 - INFO - __main__ - ***** Eval results *****
06/26/2021 18:49:30 - INFO - __main__ - acc = 0.7098795180722891
06/26/2021 18:49:30 - INFO - __main__ - auc = 0.5024916943521595
06/26/2021 18:49:30 - INFO - __main__ - f1 = 0.41516347237880497
06/26/2021 18:49:30 - INFO - __main__ - mcc = 0.0
06/26/2021 18:49:30 - INFO - __main__ - precision = 0.35493975903614455
06/26/2021 18:49:30 - INFO - __main__ - recall = 0.5
```

شکل 8.4: نتیجه تمرین به کمک 4mer، به کمک مدل DNAbert

```
06/25/2021 19:21:42 - INFO - __main__ - ***** Eval results *****
06/25/2021 19:21:42 - INFO - __main__ - acc = 0.7098795180722891
06/25/2021 19:21:42 - INFO - __main__ - auc = 0.503859617071856
06/25/2021 19:21:42 - INFO - __main__ - f1 = 0.41516347237880497
06/25/2021 19:21:42 - INFO - __main__ - mcc = 0.0
06/25/2021 19:21:42 - INFO - __main__ - precision = 0.35493975903614455
06/25/2021 19:21:42 - INFO - __main__ - recall = 0.5
```

شکل 9.4: نتیجه تمرین به کمک 6mer، به کمک مدل DNAbert

نتیجه آموزش مدل توجه

Epoch	Training Loss	Validation Loss	Accuracy
1	0.665000	0.724736	0.561959
2	0.665000	0.730071	0.561959
3	0.659300	0.699565	0.561959
4	0.652200	0.721405	0.561959
5	0.655200	0.716773	0.561959
6	0.659000	0.701253	0.561959
7	0.656900	0.733162	0.561959
8	0.650700	0.721418	0.561959
9	0.650500	0.690307	0.561959
10	0.651700	0.694987	0.561959
11	0.649500	0.724621	0.561959
12	0.650100	0.709478	0.561959
13	0.651100	0.709176	0.561959
14	0.648300	0.701109	0.561959
15	0.648600	0.723538	0.561959
16	0.651100	0.697469	0.561959
17	0.646200	0.694035	0.561959
18	0.655700	0.689684	0.561959
19	0.645500	0.708879	0.561959
20	0.646800	0.706368	0.561959

شکل 10.4: نتیجه تمرین به کمک 6mer به کمک مدل RoBerta

## فصل 5

### Discussion

دو مشکل اساسی که در داده‌ها پیدا می‌شود نويز ذاتی داده‌ها به خاطر حضور یک sgRNA در cell-line‌ها و ارگانیزم‌ها مختلف و نامتعادل بودن داده‌ها است چون معمولاً کارشناسانی که sgRNA‌های مختلف را تست می‌کنند معمولاً یک حس و ی bios از قبل روی این‌ها sgRNA و موفق بودن آنها دارند و یا به عبارتی دیگر به خاطر وقت و هزینه‌ی این آزمایش‌ها هیچ وقت ای sgRNA که فکر می‌کنند اصلاً خوب نیست را آزمایش نمی‌کنند که باعث به وجود آمدن دیتاست‌های نامتعادل می‌شود، فکر کردن راجع به راهی برای حذف این نویزها و هابios در مدل باعث می‌شود که روشی جامع برای پیش‌بینی این تاثیرگذاری هابios بدست آید.



- [1] ParsiLaTeX. <http://parsilatex.com>
- [2] Selective Breeding: [https://en.wikipedia.org/wiki/Plant\\_breeding](https://en.wikipedia.org/wiki/Plant_breeding)
- [3] TED Talk: [https://www.ted.com/talks/jennifer\\_doudna\\_we\\_can\\_now\\_edit\\_our\\_dna\\_but\\_let\\_s\\_do\\_it\\_wisely/transcript?language=fa](https://www.ted.com/talks/jennifer_doudna_we_can_now_edit_our_dna_but_let_s_do_it_wisely/transcript?language=fa)
- [4] 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. Journal of Bacteriology. 1987; 169, 5429–5433.
- [5] Mojica, F.J. , Juez, G. & Rodriguez-Valera, F. (1993) Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified PstI sites. Molecular Microbiology, 9, 613–621.
- [6] Patrick D. Hsu, Eric S. Lander, and Feng Zhang. Development and Applications of CRISPR-Cas9 for Genome Engineering Patrick. 2014, Cell 157(6): 1262-1278. 4.1.3 add gene
- [7] Opitz, D.; Maclin, R. (1999). "Popular ensemble methods: An empirical study". Journal of Artificial Intelligence Research. 11: 169–198. doi:10.1613/jair.614.
- [8] Polikar, R. (2006). "Ensemble based systems in decision making". IEEE Circuits and Systems Magazine. 6 (3): 21–45. doi:10.1109/MCAS.2006.1688199. S2CID 18032543.
- [9] :Rokach, L. (2010). "Ensemble-based classifiers". Artificial Intelligence Review. 33 (1–2): 1–39. doi:10.1007/s10462-009-9124-7. S2CID 11149239
- [10] Blockeel H. (2011). "Hypothesis Space". Encyclopedia of Machine Learning: 511–513. doi:10.1007/978-0-387-30164-8\_373. ISBN 978-0-387-30768-8.
- [11] Yann Lecun (2020). Deep Learning course at NYU, Spring 2020, video lecture Week 6. Event occurs at 53:00. Retrieved 2021-12-13.
- [12] "Hybrid computing using a neural network with dynamic external memory". Nature. 538 (7626): 471–476. 2016-10-12. Bibcode:2016Natur.538..471G. doi:10.1038/nature20101. ISSN 1476-4687. PMID 27732574. S2CID 205251479.
- [13] nature20101. ISSN 1476-4687. PMID 27732574. S2CID 205251479.
- [14] Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz; Polosukhin, Illia (2017-12-05). "Attention Is All You Need". arXiv:1706.03762 [cs.CL].
- [15] Ramachandran, Prajit; Parmar, Niki; Vaswani, Ashish; Bello, Irwan; Levskaya, Anselm; Shlens, Jonathon (2019-06-13). "Stand-Alone Self-Attention in Vision Models". arXiv:1906.05909 [cs.CV].
- [16] Jaegle, Andrew; Gimeno, Felix; Brock, Andrew; Zisserman, Andrew; Vinyals, Oriol; Carreira, Joao (2021-06-22). "Perceiver: General Perception with Iterative Attention". arXiv:2103.03206 [cs.CV].

- [17] Ray, Tiernan. "Google's Supermodel: DeepMind Perceiver is a step on the road to an AI machine that could process anything and everything". ZDNet. Retrieved 2021-08-19.
- [18] Guohui Chuai, Qi Liu et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. 2018 (Manuscript submitted)
- [19] Stemmer, M., Thumberger, T., del Sol Keyer, M., Wittbrodt, J. and Mateo, J.L. CCTop: an intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. PLOS ONE (2015). doi: 10.1371/journal.pone.0124633
- [20] Labuhn, M., Adams, F. F., Ng, M., Knoess, S., Schambach, A., Charpentier, E. M., ... Heckl, D. Refined sgRNA efficacy prediction improves large- and small-scale CRISPR–Cas9 applications. Nucleic Acids Research (2017). doi: 10.1093/nar/gkx1268
- [21] Off-target predictions in CRISPR-Cas9 gene editing using deep learning Jiecong Lin and Ka-Chun Wong, Bioinformatics. 2018 Sep 1, doi: 10.1093/bioinformatics/bty554
- [22] Guohui Chuai, Qi Liu et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. 2018 (Manuscript submitted)
- [23] CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens, Jean-Paul Concordet, Maximilian Haeussler, Nucleic Acids Research, Volume 46, Issue W1, 2 July 2018, Pages W242–W245, <https://doi.org/10.1093/nar/gky354>
- [24] E-CRISP: fast CRISPR target site identification , Florian Heigwer, Grainne Kerr & Michael Boutros
- [25] Cas-Designer: a web-based tool for choice of CRISPR-Cas9 target sites Jeongbin Park, Sangsu Bae, Jin-Soo Kim, <https://doi.org/10.1093/bioinformatics/btv537>
- [26] Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases, Sangsu Bae 1, Jeongbin Park, Jin-Soo Kim, DOI: 10.1093/bioinformatics/btu048
- [27] CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing
- [28] CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering Kornel Labun, Tessa G. Montague, James A. Gagnon, Summer B. Thyme, Eivind Valen
- [29] CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing
- [30] All you need about CRISPR: <https://www.addgene.org/guides/crispr/>
- [31] Good Overview by Wired: <https://www.wired.com/2015/07/crispr-dna-editing-2/>
- [32] DNA: <https://en.wikipedia.org/wiki/DNA>
- [33] <https://medlineplus.gov/genetics/understanding/basics/dna/>
- [34] Radiation research: <https://www.amusingplanet.com/2013/03/atomic-gardening-breeding-plants-with.html>
- [35] inserting DNA snippets into organisms: [http://www.genomenetwork.org/resources/timeline/1977\\_Gilbert.php](http://www.genomenetwork.org/resources/timeline/1977_Gilbert.php)
- [36] First genetically modified animal: <https://www.pnas.org/content/71/4/1250?tab=author-info>
- [37] First GM patent: <https://patents.google.com/patent/US4259444>

- [38] chemicals produced by GMOs: <https://pubmed.ncbi.nlm.nih.gov/6337396/>
- [39] <https://pubmed.ncbi.nlm.nih.gov/15580495/>
- [40] <https://th.schattauer.de/contents/archive/issue/721/manuscript/9641.html>
- [41] Flavr Savr Tomato: <https://calag.ucanr.edu/Archive/?article=ca.v054n04p6>
- [42] First Human Engineering: <https://www.worldscientific.com/worldscibooks/10.1142/9542>
- [43] glowing fish: <https://www.glofish.com/>
- [44] CRISPR: <http://go.nature.com/24Nhykm>
- [45] <https://en.wikipedia.org/wiki/CRISPR>
- [46] HIV cut from cells and rats with CRISPR: <https://www.nature.com/articles/531156a>
- [47] Elimination of HIV-1 Genomes from Human T-lymphoid Cells by CRISPR/Cas9 Gene Editing; Rafal Kaminski, Yilan Chen, Tracy Fischer, Ellen Tedaldi, Alessandro Napoli, Yonggang Zhang, Jonathan Karn, Wenhui Hu & Kamel Khalili
- [48] first human CRISPR trials fighting cancer: <https://time.com/4340722/hiv-removed-using-crispr/>
- [49] first human CRISPR trial approved by Chinese for August 2016 <https://www.nature.com/articles/nature.2016.20137>
- [50] <https://www.youtube.com/watch?v=jAhjPd4uNFY&t=122s>

## **Abstract**

Clustered Regularly Interspaced Short Palindromic Repeats, or in short, CRISPR is a relatively new technology that enables geneticists and medical researchers to edit parts of the genome by removing, adding, or altering parts of the DNA. Initially found in the genomes of prokaryotic organisms such as bacteria and archaea, this technology can cure many illnesses such as blindness and cancer. A significant issue for a practical application of CRISPR systems is accurately predicting the single guide RNA (sgRNA) on-target efficacy and off-target sensitivity. While some methods classify these designs, most algorithms are on separate data with different genes and cells. The lack of generalizability of these methods hinders the use of this guide in clinical trials since, for each treatment, the process must be designed with its unique dataset, which has its own problems. Here we are trying to solve the generalizability of this problem and present general and targeted prediction models that will help researchers optimize the design of sgRNAs with high sensitivity. First, we tackled the problem by leveraging Latent Profile Analysis and attention-based models to combine previous algorithms. However, the results obtained using these methods were not satisfactory since the data was noisy. Finally, we proposed a novel Ensemble Learning, which is compatible in terms of accuracy. However, our method provides the advantage of generalizability, allowing the model to offer insightful estimates to RNA on-target efficiency that can quickly learn to predict even in new genes or cells.



Sharif University of Technology  
Department of Mathematical Sciences

**M.Sc. Thesis**  
**Applied Mathematics**

# **A study in genome editing with clustered regularly interspaced short palindromic repeats**

By

Mohammad Rostami

Supervisor

Dr. Mohsen Sharifi Tabar

Second Supervisor

Dr. Hamidreza Rabiee

Advisor

Dr. Mohammad Hossein Rohban

February 9, 2022