دانشگاه صنعتی شریف
دانشکده علوم ریاضی

پایان نامه کارشناسی‌ارشد
ریاضی کاربردی

تحقیق در ویرایش ژنوم به کمک تناوب‌هایِ کوتاهِ پالیندرومِ فاصله‌دارِ منظمِ خوشه‌ای

نگارش
محمد رستمی

استاد راهنما
دکتر محسن شریفی تبار

استاد راهنمای دوم
دکتر حمیدرضا ربیعی

استاد مشاور
دکتر محمدحسین رهبان

17 بهمن 1400

به نام او

دانشگاه صنعتی شریف

دانشکده علوم ریاضی

پایان نامه کارشناسی‌ارشد

عنوان: تحقیق در ویرایش ژنوم به کمک تناوب‌هایِ کوتاهِ پالیندرومِ فاصله‌دارِ منظمِ خوشه‌اینگارش: محمد رستمی

کمیته داوران

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

تاریخ: .....

قدردانی

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

و تشکر از خداوند.

چکیده

تناوب‌هایِ کوتاهِ پالیندرومِ فاصله‌دارِ منظمِ خوشه‌ای یا به طور خلاصه ، CRISPR (کریسپر) یکی از روش‌های نسبتا نوین است که متخصصانِ ژنتیکَ و محققانِ پزشکی را قادرَ می سازد تا با حذف بخشهایی از ژنوم ، افزودن یا تغییر بخش هایی 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 روش ها این طرح ها را طبقه بندی می کنند ، بیشتر الگوریتم ها بر روی داده های جداگانه با ژن ها و سلول های مختلف هستند. عدم تعمیم این روش ها مانع استفاده از این راهنما در آزمایشات بالینی می شود ، زیرا برای هر درمان، این فرایند باید دقیقا برای همان سلول درست شده باشد و عموما 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 به‌صورت عمومی و هدفمند دقت خوبی داشته باشد تا که محققان در بهینه سازی طراح راهنمای آران‌ای با حساسیت مناسب کمک کند.

فهرست مطالب

1 مقدمات 1
1.1 RNA 1
2.1 دی‌ان‌ای 1
1.2.1 تفاوت‌های دی‌ان‌ای و آران‌ای 2
3.1 ویرایش ژنوم 3
1.3.1 شکست و تعمیر دی‌ان‌ای 3
2.3.1 Zinc finger nucleases (ZFN) 4
3.3.1 TALEN 4
4.1 CRISPR 5
1.4.1 کریسپر در باکتری 5
2.4.1 عمل کرد کریسپر در ژن 5
3.4.1 حساسیت 6
4.4.1 تاثیرگذاری 6
5.4.1 انواع کریسپر 7

2 Literature the of Review 10
1.2 روش‌های مستقیم 10
1.1.2 Chopchop 10
2.1.2 Cas-OFFinder و Cas-Designer 10
3.1.2 E-CRISP 11
4.1.2 CRISPOR 12
2.2 روش‌های یادگیری ژرف 12
1.2.2 پیش‌بینی off-target به کمک یادگیری ژرف 12
2.2.2 CCTop 13
3.2 DeepCRISPR 13
4.2 CRISPR_ONT 13

3 Methods 14
1.3 ensemble 14
1.1.3 تعریف 14
2.3 Attention 15

4 Results 16
1.0.4 Ensemble 16
2.0.4 Attention 17

5 Discussion 19

فهرست تصاویر

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. 1
2.1 شکل دو بعدی DNA 2
3.1 مقایسه دی‌ان‌ای و آران‌ای 2
4.1 مکانیزم ترمیم DNA 3
5.1 مکانیزم TALEN 4
6.1 مکانیزم ساده شده‌ای از CRISPR 5
7.1 مکانیزم CRISPR 6
8.1 مکانیزم TALEN 7

1.2 (الف) شماتیک مکان‌های off-tagets را با برآمدگی DNA یا RNA نشان می‌دهد. (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 11
3.2 هم بستگی اسپیرمن بین امتیاز موثر بودن و داده‌ها 12
4.2 هم بستگی اسپیرمن بین امتیاز موثر بودن و داده‌ها 12

1.4 ROC AUC 16
2.4 PR AUC 16
3.4 score F1 17
4.4 هم بستگی اسپیرمن بین امتیاز موثر بودن و داده‌ها 17
5.4 نتیجه تمرین به کمک ،3mer به کمک مدل DNAbert 17
6.4 نتیجه تمرین به کمک ،4mer به کمک مدل DNAbert 17
7.4 نتیجه تمرین به کمک ،6mer به کمک مدل DNAbert 18
8.4 نتیجه تمرین به کمک ،6mer به کمک مدل RoBerta 18

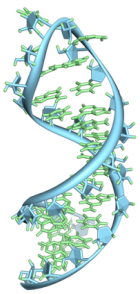
فصل 1

مقدمات

قبل از این که با کریسپر آشنا شویم، خوب است کمی راجع به تاریخچه ویرایش ژن‌ها صحبت کنیم. انسان‌ها سال‌هاست که مشغول به ویرایش و مهندسی ژن هستند، با استفاده از پرورش انتخابی[1]. اصلاحات نژادی متعددی در گیاهان و حیوانات مخصوصا گونه‌های کلیدی مانند گندم، برنج و سگ‌ها ایجاد شده است. انسان‌ها در این کار شدیدا ماهر شدن به‌طوری که در صده گذشته، تعداد دانه‌های هر شاخه گندم چندین برابر و ارتفاع آنها کوتاه‌تر شده تا در معرض خطر کمتری باشند و حدود ۸۰ نژاد جدید سگ به وجود آمده است. البته با وجود پیشرفت‌های متعدد انسان‌ها تا کشف دی‌ان‌ای دقیقا ساز و کار آن را نمی دانستند.

1.1 RNA

اسید ریبونوکلئیک[2] یا RNA یک مولکول پلیمری است که در نقش‌های بیولوژیکی مختلف مانند کدگذاری، رمزگشایی، تنظیم و بیان ژن‌ها ضروری است. RNA به صورت یک رشته منفرد از نوکلئوتیدها (بازهای نیتروژنی گوانین، اوراسیل، آدنین و سیتوزین که با حروف G، U A و C مشخص می شوند) است که برخودش تا می خورد، بر خلاف دی‌ان‌ای که با یک رشته دیگر جفت شده است.



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.

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

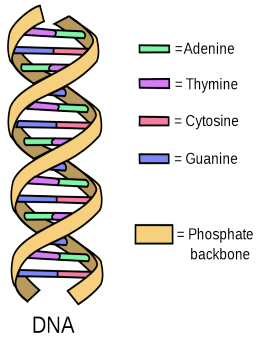
2.1 دی‌ان‌ای

دئوکسی ریبو نوکلئیک اسید[3] به اختصار دی‌اِن‌ای یک مولکول متشکل از دو زنجیره پلی نوکلئوتیدی است که به دور یکدیگر می‌پیچند تا دستورالعمل‌های ژنتیکی برای کارکرد و توسعهٔ زیستی جانداران و ویروس‌ها مورد استفاده قرار می‌گیرد. نقش اصلی مولکول دی‌ان‌ای ذخیره‌سازی طولانی مدت اطلاعات ژنتیکی و دستوری است. لیپیدها، پروتئین‌ها، کربوهیدرات های پیچیده (پلی ساکاریدها) و اسیدهای نوکلئیک سه درشت‌مولکول‌های 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 ساده‌تری به نام نوکلئوتید تشکیل شده اند. هر نوکلئوتید از یکی از چهار نوکلئوباز حاوی نیتروژن (سیتوزین C، گوانین G، آدنین A یا تیمین T)، کربوهیدرات پنج‌کربنه به نام دئوکسی

[1] Selective Breeding
[2] RiboNucleic Acid
[3] 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 می شود. 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 می شوند 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 می شوند. در DNA، پیریمیدین ها 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 می شود. بخش بزرگی از DNA (بیش از ٪98 برای انسان) بی‌کد[4] 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 حلقهٔ شکری نوکلئوتید شماره‌گذاری شده‌اند. هر رشتهٔ دی‌ان‌ای یا آران‌ای دارای یک پایانهٔ ′۵ که معمولا شامل یک گروه فسفاتی است و یک پایانهٔ ′۳ 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 می کند. 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.

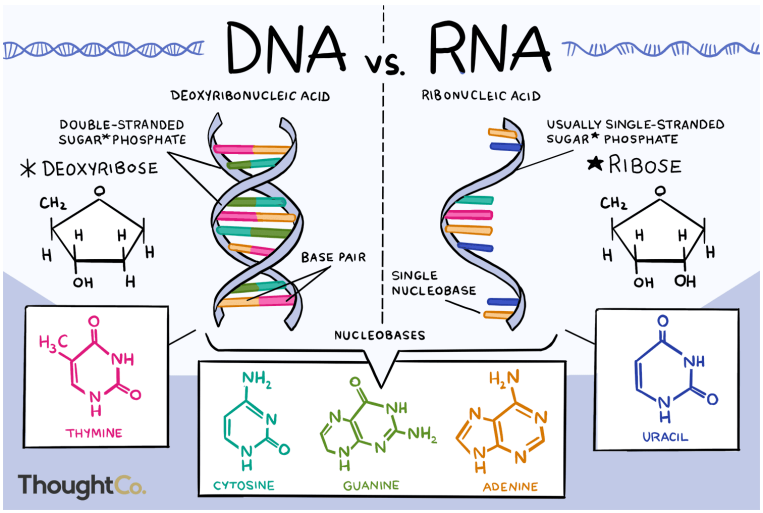
1.2.1 تفاوت‌های دی‌ان‌ای و آران‌ای

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 آلی یوراسیل و در RNA باز آلی تیمین شرکت ندارد (U در DNA و T در RNA).
- قند پنج‌کربنه موجود در DNA را دئوکسی ریبوز و در RNA قند ریبوز نامیده می شود. تفاوت بین قندها وجود گروه هیدروکسیل بر روی کربن ′۲ ریبوز و عدم وجود آن در کربن ′۲ دئوکسی ریبوز است.
- DNA برعکس RNA از هستهٔ 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

[4] 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 رادیواکتویی بر روی گیاهان دست به تغییر ژنوم آنها به طور کاملا 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.

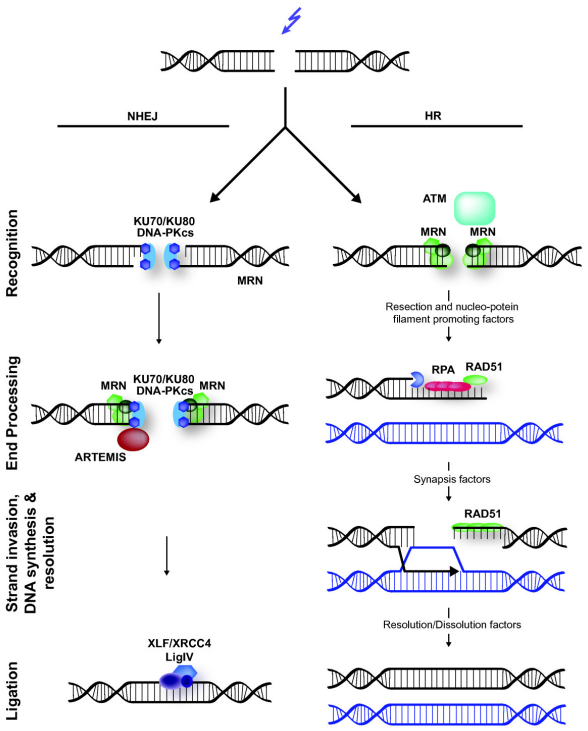
تا کنون سه تکنیک موفق و معروف برای ویرایش ژنوم مهندسی شده است: نوکلئاز انگشت روی[5] (ZFNs) ، نوکلئازهای اثرگذار شبه فعال کننده رونویسی[6] (TALEN) ، و سیستم تناوب‌هایِ کوتاهِ پالیندرومِ فاصله‌دارِ منظمِ خوشه‌ای[7] (CRISPR). کلید ویرایش ژنوم ایجاد شکست دو رشته‌ی DNA در نقطه مورد نظر است و این سه روش مبتنی بر شکستِ 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[8] (DSB) تکیه دارد. دو مسیر اصلی وجود دارد که DSB را تعمیر می‌کند. اتصال انتهای غیر همولوگ[9] (NHEJ) و تعمیر هدایت شده همولوژی[10] (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 می توان با ایجاد یک بردار با عناصر ژنتیکی مورد نظر در یک توالی که همولوگ با توالی های کناری یک DSB است مورد استفاده قرار داد. این باعث می شود 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 رشته ی DNA می شوند .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 می رود، زیرا همواره چندین نوکلئوتید در جایگاه ترمیم از بین می روند و دو انتهای شکسته شده از کروموزوم های همولوگ یا غیر همولوگ به یکدیگر متصل می شوند. زمانی که کروماتیدهای خواهری برای ترمیم شکست های دو رشته ای 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 می یابند، آن گاه در هر انتها چندین باز توسط نوکئاز حذف شده و دو مولکول از طریق آنزیم لیگاز به هم متصل می گردند. DSBها ترجیحاً در سلول توسط اتصال انتهایی غیر همولوگ (NHEJ) ترمیم می شوند، مکانیزم سریعی که اغلب باعث درج یا حذف (indels) در DNA می شود. ایندل ها اغلب منجر به تغییر اساسی در DNA می شوند و به‌طوری که DNA عملکرد خود را از دست می دهند. پس در نتیجه معمولا به عنوان سلول مرده درنظر 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 DBS، به این روش تعمیر نشود.



شکل 4.1: مکانیزم ترمیم DNA

[5]Zinc Finger Nuclease
[6]Transcription activator-like effector nuclease
[7]Clustered Regularly Interspaced Short Palindromic Repeats
[8]Double-Strand Break (Cut)
[9]Non-Homologous End Joining
[10]Homology Directed Repair

HDR

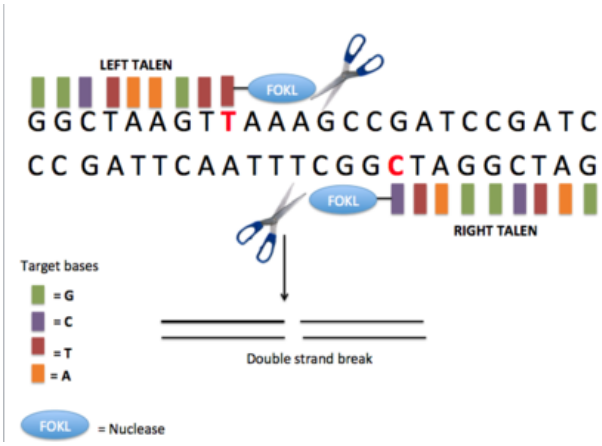
تعمیر هدایت شده همولوژی (HDR) مکانیزمی در سلول ها
برای ترمیم ضایعات DNA دو رشته‌ای از یک نسخه مشابه DNA برای ترمیم استفاده می کند. رایج ترین شکل HDR نوترکیبی همولوگ است. مکانیسم HDR تنها زمانی می تواند توسط سلول استفاده شود که یک قطعه همولوگ از DNA در هسته وجود داشته باشد، عمدتاً در فاز G2 و S چرخه سلولی. 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.

2.3.1 Zinc finger nucleases (ZFN)

نوکلئاز انگشت روی یا ZFN اولین سیستم پروتئینی متصل شونده به DNA قابل برنامه ریزی با کاربرد وسیع است.ZFNها شامل زنجیره ای از پروتئین های انگشت روی هستند که به یک نوکلئاز باکتریایی ملحق شده اند تا بتوانند سیستمی را تولید کنند که قادر به ایجاد برش های دو رشته ای خاص در DNA برای ویرایش ژن باشد. پروتئین های انگشت روی هدف قرار دادن ناحیه خاص را فراهم می کنند زیرا هر یک از آنها سه جفت باز یا ۳bp از DNA را شناسایی می کنند. نوکلئازی که معمولاً در تکنولوژی ZFN متصل به زنجیره پروتئین های انگشت روی است FokI نام دارد که برای اتصال به DNA باید دایمریزه شده باشد، بنابراین یک جفت از ZFN برای هدف گیری و برش DNA مورد استفاده قرار می گیرد.این آنزیم ها کمک زیادی به تولید موجودات ترانسژنیک می کنند و بدلیل اینکه فراوانی نوترکیبی همولوگ بسیار ناچیز بوده اهمیت زیادی در مهندسی ژنتیک و مطالعات ترانسژنیک ، ناک اوت و غیره پیدا کرده اند. این پروتئین های مهندسی شده متصل شونده به DNA می توانند ژنوم را در جایگاه های ویژه ای شناسایی کرده و ایجاد برش های دورشته ای کنند . در صورتیکه سیستم تعمیر NHEJ فعال شود چون این سیستم ترمیم مستعد خطاست سبب ایجاد جهش در آن ناحیه خاص از ژنوم می شود بنابراین در مطالعات موتاژنز نیز اهمیت دارند. انتقال یک وکتور حاوی ژن مورد نظر به همراه ZFNs سبب تسهیل درج ژن در آن ناحیه از ژنوم می گردد.

3.3.1 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 آنزیم های محدودکننده مصنوعی هستند که با ادغام حوزه برش DNA یک نوکلئاز با دامنه های TALE طراحی شده اند، که می توانند به طور خاص یک توالی DNA منحصر به فرد را شناسایی کنند. این پروتئین‌های ادغام شده به‌عنوان قیچی DNA به‌راحتی 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 موثر یک رشته ۳۴تایی 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 می تواند با اتصال انتهایی غیر همولوگ (NHEJ) ترمیم شود، که منجر به اختلالات ژنی از طریق وارد کردن یا حذف های کوچک می شود. هر تکرار حفظ می شود، به جز آمینواسید ۱۳ و ۱۲ که به آنها دو باقیمانده متغیر تکرار (RVDs) می‌گویند. RVD ها توالی DNA را تعیین می کنند که TALE به آن متصل می شود. 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 می توان با کاتالیزوری از یک نوکلئاز از DNA به نام FokI، ادغام کرد تا با آن هاTALEN را ساخت. ساختارهای TALEN توالی‌های DNA را فقط در مکان‌های از پیش انتخاب شده متصل می‌کنند و می‌شکنند. هدف TALEN را می توان بر اساس یک کد آسان پیش بینی کرد. نوکلئازهای TAL تا حدی به دلیل طولشان که بیش از 30 جفت است میتوان فقط مختص آن هدف درنظر گرفت. TALEN را می توان در محدوده 6 جفت باز از هر نوکلئوتید منفرد در کل ژنوم انجام داد.

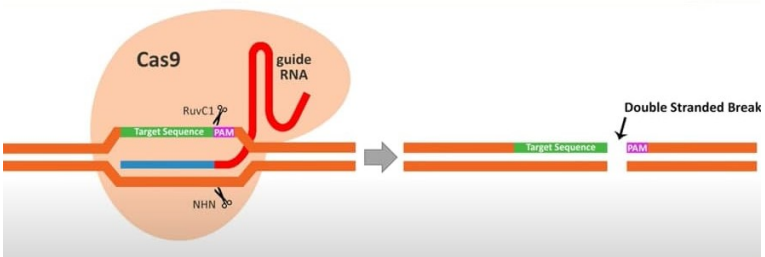


شکل 5.1: مکانیزم TALEN

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

4.1 CRISPR

CRISPR (به معنی "تناوب‌هایِ کوتاهِ پالیندرومِ فاصله‌دارِ منظمِ خوشه‌ای" Clustered Regularly Interspaced Short Palindromic Repeats) یا به اختصار کریسپر (به انگلیسی: بخشی از دی‌ان‌ای پروکاریوت هستند که حاوی تناوب‌های کوتاهِ توالی‌های بنیادین هستند. بخشی از سیستم کریسپر "پروتئین Cas9" است. این پروتئین قابلیت جستجو، برش زدن و تغییر دی ان ای(DNA) را دارد. قبل از این تکنیک از روش "تحویل یا انتقال ژن" استفاده می‌شد، به این صورت که از یک ناقل ویروسی یا غیرویروسی 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 "کَس9" (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

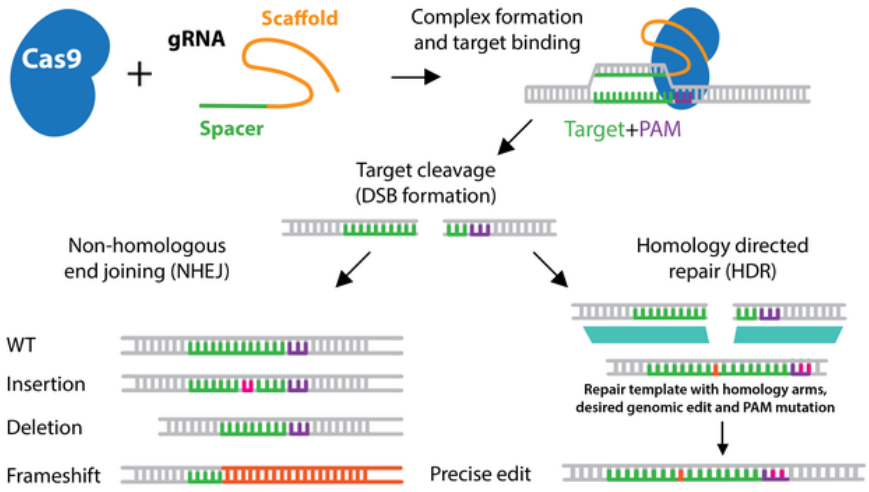
1.4.1 کریسپر در باکتری

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. آرایهٔ 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 ریبونوکلئازهای 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 RNaseهایی که هنوز مشخص نشده‌اند برای تولید رونوشت کاملاً بالغ پردازش می‌شوند. CRISPR نوعII، یک RNA کریسپر فعال‌کننده 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. هایCrRNA 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 نوعI، کمپلکس Cascade با یک مولکول CrRNA لود می‌شود که یک مجموعه نظارتی بی نظیری است که DNA هدف را شناسایی می‌کند. سپس نوکلئازCas3، لوپ Cascade R را به کار گرفته و به آن متصل می‌شود و واسطه تخریب توالی هدف می‌شود. در CRISPR نوع III, هاCrRNA یا به کمپلکس‌های Csm یا به کمپلکس‌های Cmr به ترتیب متصل شده و به ترتیب سوبستراهای DNA و RNA را می‌شکنند. درمقابل، 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 عمل کرد کریسپر در ژن

همانطور که گفتیم، مدل‌های مختلفی از CRISPR تا به حال درست شده است 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 دقیقا 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 کاملا 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 که هر یک ویژگی‌های خاص خود را دارند ولی ما در پژوهش خود ساده‌ترین مورد آن یعنی cas9 به همراه یک RNA که به آن sgRNA یا single guide RNA می گویند، 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 نوکلیوتید 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 هدف گیری شده توسط gRNA در مقایسه با بقیه ژنوم تعیین می شود. 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 دقیقا هدف را ویرایش می‌دهد نا 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 می شوند و باید هنگام طراحی یک gRNA برای آزمایش خود در نظر گرفته شوند.

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

بنابراین، دو نیکاز که رشته‌های DNA مخالف را هدف قرار می دهند 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 می تواند با ویرایش ژن با واسطه HDR برای ویرایش های ژنی خاص ترکیب شود.

در سال 2015، محققان از rational mutagenesis برای توسعه دو Cas9 با ثبات بالا استفاده کردند: eSpCas9 و SpCas9-HF1. eSpCas9 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 می دهد که برهمکنش Cas9 با ستون فقرات فسفات DNA را مختل می کند. 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 می کنند.

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 دآمیناز مانند APOBEC ایجاد می‌شوند. ویرایشگرهای پایه توسط یک gRNA به یک مکان خاص قرار می‌گیرند و می توانند 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 دآمینازها در طبیعت وجود ندارند، اما با تکامل هدایت شده Escherichia coli TadA، یک tRNA آدنین دآمیناز ایجاد شده اند. مانند ویرایشگرهای پایه سیتوزین، دامنه تکامل یافته TadA با پروتئین Cas9 ترکیب می شود 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 دآمیناز 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.

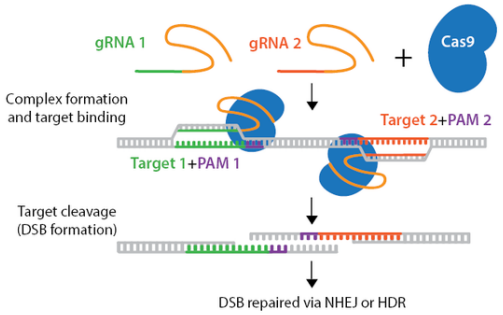
5.4.1 انواع کریسپر

طبقه‌بندیrova 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 خارجیDNA 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. برنامه های کاربردی. نمونه‌هایی از این سیستم‌های ویرایش پایه عبارتند از جهش‌زایی هدفمند با واسطه AID (TAM) و CRISPR-X، که در آن Cas9 با سیتیدین دآمیناز (AID) ناشی از فعال‌سازی ترکیب می‌شود.



شکل 8.1: مکانیزم TALEN

نیکاز CRISPR/Cas جهش‌یافته، به‌جای 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 رشته ای (DSB) ایجاد می کنند که با استفاده از اتصال انتهایی غیر همولوگ (NHEJ) و مستعد خطا تعمیر می شود. 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 S. pyogenes Cas9 (SpCas9) مطمئناً متداول ترین اندونوکلئاز CRISPR برای مهندسی ژنوم است، ممکن است بهترین اندونوکلئاز 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، نگران کننده 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 توالی های PAM غیر NGG شد. جایگزین دیگر Cas9، xCas9 است که مجموعه وسیعی از توالی های PAM مانند GAA، NG و GAT را هدف قرار می دهد، در حالی که حداقل فعالیت خارج از هدف را نیز نشان می دهد.

اصطلاح	تعریف
ویرایشگر پایه (Base editor)	ادغام یک پروتئین Cas به یک دآمیناز که تبدیل مستقیم باز در RNA یا DNA را بدون شکست دو رشته DNA امکان پذیر می کند.
Cas	CRISPR Associated Protein, شامل نوکلئازهایی مانند Cas9 و Cas12a (همچنین به عنوان Cpf1 شناخته می شود)
CRISPR	تناوب‌هایِ کوتاهِ پالیندرومِ فاصله‌دارِ منظمِ 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 2 gRNA 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 scafold 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*
Streptococcus pyogenes (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
Staphylococcus aureus (SA); SaCas9	3′ NNGRRT or NNGRR(N)
Acidaminococcus sp. (AsCpf1) and Lachnospiraceae bacterium (LbCpf1)	5′ TTTV
AsCpf1 RR variant	5′ TYCV
LbCpf1 RR variant	5′ TYCV
AsCpf1 RVR variant	5′ TATV
Campylobacter jejuni (CJ)	3′ NNNNRYAC
Neisseria meningitidis (NM)	3′ NNNNGATT
Streptococcus thermophilus (ST)	3′ NNAGAAW
Treponema denticola (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 می کنیم، 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

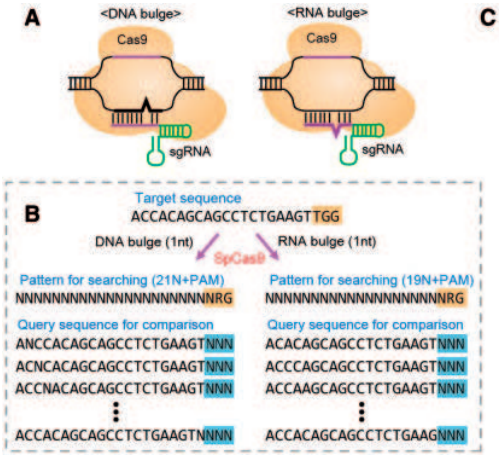
1.1.2 Chopchop

این مقاله که الگوریتم خود را سه بار بروزرسانی کرده است، به عنوان ورودی رشته DNA ورودی و یا اسم ژن یا مختسات آن را می گیرد هم چنین مورد استفاده‌ی طرح را می پرسد. 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 می کنند و primer3 برای پیدا کردن primer ها استفاده می‌کند، این الگوریتم با توجه به پژوهش‌های قبلی از ٦ ویژگی مهم برای مرتب کردن طرح‌ها استفاده می‌کنند که عبارت اند از: تعداد offtarget ها، معماری ژن، GC-Content، وجود نوکلوید G در ۲۰امین 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 درسترس 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 نوکلیوتایدها 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 می شود، 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 می کند و با توجه به آنها امیتاز حساسیت می دهد.

2.1.2 Cas-OFFinder و Cas-Designer

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 ازCRISPR/Cas نوع ،II که در حال حاضر به طور گسترده برای تحقیقات زیست پزشکی و بیوتکنولوژی استفاده می‌شود. Cas-Designer به سرعت ارائه می دهد فهرستی از تمام توالی های 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 می دهد تا به کاربران کمک کند سایت های مناسب برای ژن را برای Knockout انتخاب کنند. Cas-Designer نتایج را در یک جدول تعاملی نشان می دهد و فیلتر کاربر پسند را ارائه می دهد کارکرد.

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



C

RGEN Target (5' to 3')	Position	Cleavage Position (%)	Direction	GC Contents (%, w/o PAM)	Out-of-frame Score	Mismatches 0	1	2	DNA Bulge	RNA Bulge
TAGGACCAATAAGTCTTAATTGG	68	41.1	+	30.0	67.5	1	0	0	Show	Show
TCTGCAGGTTCAGAATTATAGGG	118	59.9	-	35.0	33.2	1	0	9	Show	Show
TGTGGAGTTTTAAATAGGTTTGG	172	86.0	-	30.0	71.2	1	0	1	Show	Show
ACCTATTTAAAACTCCACAAAGG	176	93.2	+	30.0	68.9	1	0	2	Show	Show

Bulge Type	Target	Chromosome	Position	Direction	Mismatches	Bulge Size	Info
DNA	crRNA: ACC-TATTTAAAACTCCACAANGG DNA: ACCATAgTgAAAAACTCCACAAGGG	chr12	13498156	-	2	1	Info at Ensembl
DNA	crRNA: ACCTATTTAAAACTCCACA-ANGG DNA: ACCTATTTAAAAaTCCtCAGAGGG	chr7	30312277	-	2	1	Info at Ensembl

⋮

Bulge Type	Target	Chromosome	Position	Direction	Mismatches	Bulge Size	Info
RNA	crRNA: ACCTATTTAAAACTCCACAANGG DNA: ACCTATaTAAAA-TCCAaATGG	chr8	128284008	-	2	1	Info at Ensembl
RNA	crRNA: ACCTATTTAAAACTCCACAANGG DNA: ACCTATTgAAAAtTCC-CAATGG	chr8	103798023	+	2	1	Info at Ensembl

⋮

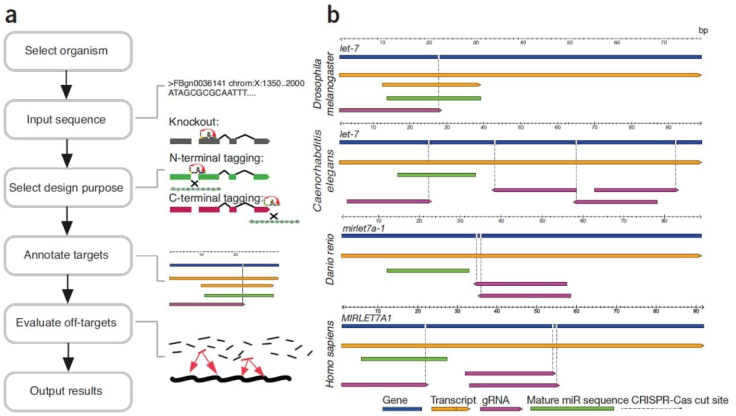
شکل 1.2: (الف) شماتیک مکان‌های off-tagets را با برآمدگی DNA یا RNA نشان می‌دهد. (ب) استراتژی برای برآمدگی 1-nt DNA یا RNA بر اساس .Cas-OFFinder ج) یک مثال از یک جدول خروجی Cas-Designer تمام gRNA های ممکن را از توالی های ورودی به همراه اطلاعات مفید (بالا) نشان می دهد. اگر کاربر روی رنگ آبی کلیک کند عدد، 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"

میکروهومولوژی را به سرعت محاسبه می کندکه با فراوانی جهش های تغییر قاب همبستگی مثبت دارد (Bae et al., 2014b). محتوای GC و امتیازات خارج از کادر در این مرحله موقعیت های برش را نشان می دهد.

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

3.1.2 E-CRISP

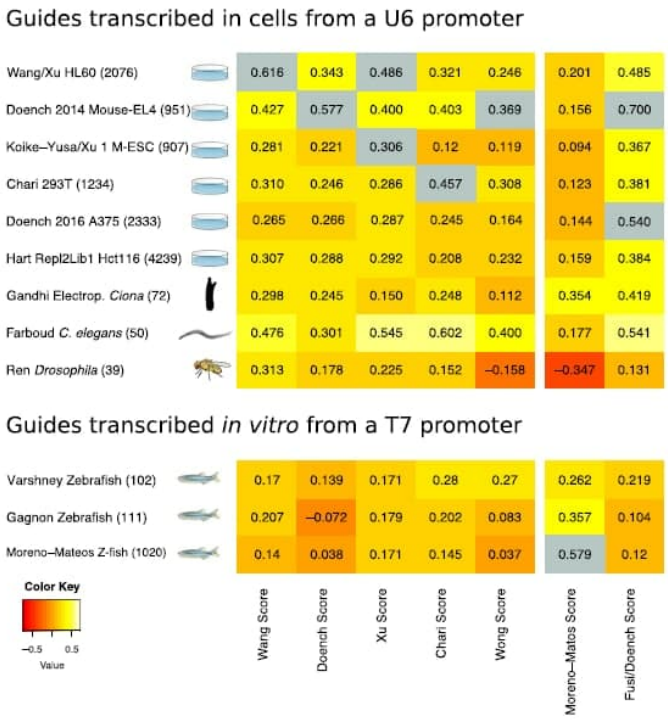
در اینجا ما E-CRISP، یک برنامه وب برای طراحی توالی های gRNA را توصیف می کنیم. (الف) مراحل .E-CRISP ابتدا کاربر ارگانیسم و دنباله هدف را انتخاب می کند. این هدف می تواند یک نماد ژن، یک شناسه ENSEMBL یا یک توالی FASTA باشد. دوم، کاربر هدف آزمایش ویرایش را مشخص می کند. بسته به هدف، E-CRISP مناطق مختلفی از توالی ژن را مورد هدف قرار می دهد. سوم، E-CRISP نتایج را با توجه به اطلاعات حاشیه نویسی ژن فیلتر می کند. چهارم، اهداف خارج از هدف بر اساس تراز توالی هر طرح با ژنوم مرجع تجزیه و تحلیل می شوند. در نهایت، E-CRISP یک صفحه خروجی تعریف شده توسط کاربر تولید می کند. (ب) RNA های راهنما در برابر جایگاه let-7 گونه های مشخص شده طراحی شده اند. توالی و محل gRNA های بالغ از miRBase بازیابی 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-CRISP توالی‌های هدف مکمل gRNA را شناسایی می‌کند که به یک موتیف که از سمت ۳′ مجاور به N(G یا A)G 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 می کند. با استفاده از این الگوریتم‌ها، می‌توان در چند ساعت کتابخانه‌هایی در مقیاس ژنومی برای چندین موجود زنده ایجاد کرد.



شکل 2.2: الگوریتم E-CRISP

4.1.2 CRISPOR

CRISPOR وب‌سایتی است که به انتخاب و بیان توالی‌های راهنمای CRISPR کمک می‌کند، که در دو مقاله توضیح داده شده است (Gen Biol 2016 و NAR 2018). در حالت پیش فرض، کاربر یک توالی DNA ورودی را چسبانده و ژنوم را انتخاب می کند. سپس CRISPOR راهنماها را در توالی ورودی فهرست می‌کند و اطلاعات مربوط به آنها را که در پایگاه‌های اطلاعاتی و الگوریتم‌ها یافت می‌شود، از جمله انواع ژنوم، امتیازهای پیش‌بینی‌شده off-targets و هدف، اضافه می‌کند. برای هر دنباله راهنما، پرایمرهای مختلفی طراحی شده است، به عنوان مثال. برای تقویت هدف، RNA های راهنما را با رونویسی آزمایشگاهی پس از بازپخت پرایمرهای همپوشانی یا برای شبیه سازی در پلاسمیدهای AddGene تولید کنید. برای پیش‌بینی، داده‌ها را از هشت مطالعه off-target، SpCas9 جمع‌آوری کرده و آنها را با سایت‌های پیش‌بینی‌شده توسط الگوریتم‌های محبوب مقایسه کردند و دریافتند که پیش‌بینی‌های off-target مبتنی بر توالی بسیار قابل اعتماد هستند، و اکثر اهداف خارج از هدف را با نرخ جهش بالاتر از 0.1٪ شناسایی می‌کنند، در حالی که تعداد موارد مثبت کاذب را می‌توان تا حد زیادی با یک برش روی این امتیاز حساسیت را افزایش داد. با توجه به آزمایشات مقاله به این درس یافتند که امتیاز موثر بودن به شدت به این بستگی دارد که آیا RNA راهنما از یک پروموتر U6 بیان می‌شود یا در شرایط آزمایشگاهی رونویسی می‌شود و با این ویژگی نشان دادند که میتوان با زمان مناسب پیش‌بینی مناسبی ارائه داد.

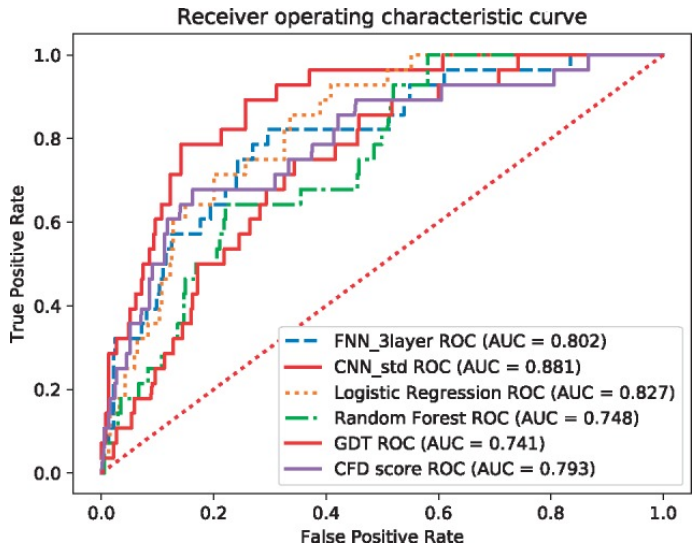


شکل 3.2: هم بستگی اسپیرمن بین امتیاز موثر بودن و داده‌ها

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. روش های پیش بینی 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 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 CFD، MIT- CROP، IT، و CCTop) و سه مدل سنتی یادگیری ماشین (یعنی جنگل تصادفی، درخت‌های تقویت‌کننده گرادیان، و رگرسیون لجستیک) در هر دو مجموعه داده از نظر مقادیر AUC، نشان دهنده لبه های رقابتی الگوریتم های پیشنهادی است. تحلیل های اضافی برای بررسی



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

دلایل زمینه ای از دیدگاه های مختلف انجام می شود.

2.2.2 CCTop

این روش برای اینکه طرح‌های مختلف که به صورت N20NGG هستند را دسته بندی 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 می کردند و ادامه آن پیشترف 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-ribosmal 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 طرخ جایگاه هر نیکلوتاید را در هایsgrna موثر و ناموثر برسی 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 طرح ها باشند و هاY امتیاز آن‌ها باشد داریم:
پس از تمرین این مدل به ۲۸ ویژگی رسید که بیشتر ویژگی ها در ناحیه اسپیسر واقع شده اند و بعضی از آنها قبلا پیدا شده بود و بعضی جدید بود:

- هاG به شدت در موقعیت های -1 و -2 به PAM در CAS9 ترجیح داده می شوند
- هاT در چهار موقعیت نزدیک به PAM نامطلوب هستند
- نوکلئوتیدهای از ۵′ به ۳′، در حالی که توالی در سمت برعکس تاثیر قابل توجهی ندارد.
- هاC در موقعیت -3 در CAS9 ترجیح داده می شوند
- هاA در موقعیت -5 تا -12 ترجیح داده می شود.
- G ها در موقعیت های -14 تا -17 ترجیح داده می شوند.

3.2 DeepCRISPR

4.2 CRISPR_ONT

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.[1][2][3] بر خلاف ensemble آماری، که معمولاً از بی نهایت 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

الگوریتم های یادگیری نظارت شده وظیفه جستجو در فضای فرضیه را برای یافتن یک فرضیه مناسب انجام می دهند که پیش بینی های خوبی را با یک مسئله خاص انجام دهد.[4]

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 الگوریتم های 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 دقیقا 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 رتبه بندی‌ها استفاده کردیم، مانند وزن دهی بر حسب دقت هر مدل روی یک دیتا ثابت و همین طور روش LPA یا Latent Profie Analysis که به ما مدلی برحسب پیشبینی مدل‌هایی دیگر می دهند. از این روش‌ها ما دو مدل بدست آوردیم ولی دقت این مدل‌ها همگی از مدل DeepCRISPR پایین تر بودند با آنالیز بیشتر به این نتیجه رسیدیم که این مدل‌ها بر سر بعضی نقاط شدیدا 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 لزوما 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 می کرد، چهار الگوریتم، ExtraTreeRegressor، GradientBoostingRegressor، LinearRegression، RandomForestRegressor را برای ensemble اکسپرت‌ها انتخاب کردیم. هر کدام از الگوریتم‌ها به تنهایی به داده آموزش حساس بودند و با انجام cross-validation لزوما به نتیجه بهتری نمی رسیدند 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 Cas-OFFinder، CCTop، E-Crisp، CRISPOR

و Chopchop جمع آوری کردیم، که منجر بدست آمدن ۵۰هزار sgRNA یکتا شد و از آنجا که خروجی الگوریتم‌ها می توانست NaN هم باشد، با حذف این داده‌ها به ۳۶هزار 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 بر خلاف 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 می کند. 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 داده ها مهم تر از سایرین است بستگی به زمینه دارد و با نزول گرادیان آموزش داده می شود.

مکانیسم‌های مانند توجه در دهه 1990 با نام هایی مانند ماژول های ضربی، واحدهای سیگما پی و ابرشبکه ها معرفی شدند. [1] انعطاف پذیری آن ناشی از نقش آن به عنوان "وزن نرم" است که می تواند در طول زمان اجرا تغییر کند، برخلاف وزنه های استاندارد که باید در زمان اجرا ثابت بمانند. کاربردهای توجه شامل حافظه در ماشین‌های تورینگ عصبی، وظایف استدلال در رایانه‌های عصبی متمایز [2]، پردازش زبان در ترانسفورماتورها، و پردازش داده‌های چندحسی (صدا، تصاویر، ویدئو، متن) در درک‌کننده‌ها است. [3] [4] [5] [6]

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 هاsgRNA 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 تاییk را به عنوان دیکشنری کلمات در نظر می گیریم. 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

1.0.4 Ensemble

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: SpearmanrResult(correlation=0.5357974645093228, pvalue=7.58034824820603e-12)
DeepCRISPR: SpearmanrResult(correlation=0.5336054772473561, pvalue=9.56026407093448e-12)
```

	sgRNA_number	KO_reporter_assay	DeepCRISPR_score	CRISPRater_score	SSC_Score	sgRNA_Scorer_score	sgRNA_Designer_rsII_score	sgRNA_sequence	extended_spacer	reg
0	sg1	0.000	0.177065	0.5710	-0.485	30.66	0.571	GAGTCGGGGTTTCGTCATGTTGG	AGTAGAGTCGGGGTTTCGTCATGTTGGTCA	0.397722
1	sg2	0.000	0.055157	0.6998	-0.266	54.96	0.533	CGCCGCCGCTTTCGGTGATGAGG	CTGCCGCCGCCGCTTTCGGTGATGAGGAAA	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	TGGGCGGATCACTTGACGTCAGG	GAGGTGGGCGGATCACTTGACGTCAGGAGT	0.175252
4	sg5	0.000	0.120955	0.6796	0.067	12.44	0.485	TTACCATAGTGTACGGGTGCAGG	CTTTTTACCATAGTGTACGGGTGCAGGCAT	0.039664
...	...	...	...	...	...	...	...	...	...	...
420	sg426	0.953	0.545577	0.7671	0.879	69.61	0.670	GCGTTGGGTGGTACTTGCAGAGG	TTGAGCGTTGGGTGGTACTTGCAGAGGTGG	0.771596
421	sg427	0.955	0.493218	0.6749	-0.154	13.28	0.555	ATGTAGGGCTTGGCGAGTCTAGG	GGATATGTAGGGCTTGGCGAGTCTAGGTCA	0.864814
422	sg428	0.955	0.568641	0.7716	0.743	93.33	0.604	GGTAGAAGGTGTAACTCCGGTGG	TCGTGGTAGAAGGTGTAACTCCGGTGGGAG	0.913264
423	sg429	0.963	0.173204	0.6069	-0.025	60.36	0.609	GTTTAGCCAAGTATCATGCATGG	AACAGTTTAGCCAAGTATCATGCATGGTTC	0.816500
424	sg430	0.973	0.409570	0.7093	0.801	92.17	0.732	GCGCGGTAGTCGTGACCCGGCGG	AGTGGCGCGGTAGTCGTGACCCGGCGGTCC	0.851474

425 rows × 10 columns

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

2.0.4 Attention

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 5.4: نتیجه تمرین به کمک 3mer، به کمک مدل DNAbert

```
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 6.4: 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 7.4: 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 8.4: نتیجه تمرین به کمک ،6mer به کمک مدل RoBerta

فصل 5

Discussion

دو مشکل اساسی که در داده‌ها پیدا می‌شود نویز ذاتی داده‌ها به خاطر حضور یک sgRNA در هاcell-line و ارگانیزم‌ها مختلف و نامتعادل بودن داده‌ها است چون معمولا کارشناسانی که sgRNA های مختلف را تست می‌کنند معمولا یک حس و یbios از قبل روی این هاsgRNA و موفق بودن آنها دارند و یا به عبارتی دیگر به خاطر وقت و هزینه‌ی این آزمایش‌ها هیچ وقت ایsgRNA که فکر میکنند اصلا خوب نیست را آزمایش نمی‌کنند که باعث به وجود آمدن دیتاست‌های نامتعادل می‌شود، فکر کردن راجع به راهی برای حذف این نویزها و هاbios در مدل باعث می‌شود که روشی جامع برای پیش‌بینی این تاثیرگذاری هاsgRNA بدست آید.

مراجع

[1] ParsiLaTeX. http://parsilatex.com

[2] Selective Breeding: https://en.wikipedia.org/wiki/Plant_breeding

[3] یاداشت‌های کلاس

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 Ensemble Learning techniques to combine previous algorithms. However, the results obtained using these methods were not satisfactory since they had a considerable disagreement. Finally, we proposed a novel attention-based model, 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 6, 2022