دانشگاه صنعتی شریف
دانشکده علوم ریاضی

پایان نامه کارشناسی‌ارشد
ریاضی کاربردی

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

نگارش
محمد رستمی

استاد راهنما
دکتر محسن شریفی تبار

استاد راهنمای دوم
دکتر حمیدرضا ربیعی

استاد مشاور
دکتر محمدحسین رهبان

11 فروردین 1401

به نام او

دانشگاه صنعتی شریف

دانشکده علوم ریاضی

پایان نامه کارشناسی‌ارشد

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

نگارش: 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 پالیندروم فاصله‌دارِ منظمِ خوشه‌ای یا به طور خلاصه، کریسپر (CRISPR) یکی از روش‌های نسبتا نوین است که متخصصانِ ژنتیک و محققانِ پزشکی را قادر می سازد تا با حذف بخشهایی از ژنوم ، افزودن یا تغییر بخش هایی از آن در دی‌ان‌ای (DNA) 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، پیش‌بینی دقیق تاثیر راهنمای آران‌ای (Guide RNA) 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. در این پژوهش روشی پایدار برای ادغام نتایج روش‌های مختلف برای تخمین دقیق تأثیرگذاری یک راهنما ارائه می‌دهیم. از آنجایی که این روش با تعداد داده کمی دارای دقت بالایی است روشی مناسب برای استفاده در مسئله‌هایی است که تعداد داده بسیار کم است.

فهرست مطالب

فهرست تصاویر

فصل 1

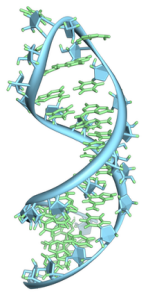
مقدمه

مقیاس در حال گسترش و پیچیدگی ذاتی داده‌های بیولوژیکی، استفاده روزافزون از یادگیری ماشین در زیست‌شناسی را برای ساختن مدل‌های آموزنده و پیش‌بینی‌کننده فرآیندهای بیولوژیکی اساسی تشویق کرده است. در ویرایش ژن‌ها نیز این روش‌ها موثر هستند زیرا تعداد عوامل موثر در موفقیت ویرایش ژن (تأثیرگذاری) بسیار بالا و نقش هر کدام از ویژگی‌ها مبهم است، علاوه بر آن بدست آوردن تمام این عوامل پیچیده و هزینه بر است و همچین پیش‌بینی اثرات بوجود آمده و مناطق تغییر کرده ناخواسته (حساسیت) کاری سخت و تصادفی است که برای مدل‌های یادگیری ماشین امر مرسوم است. عموما روش‌های ویرایش ژن‌ها امری هزینه بر با داده‌های کم است ولی با پیشرفت علم روشی مناسب و کم هزینه به نام کریسپر برای ویرایش ژن بدست آمده است ولی 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 شده‌اند 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 نوکلئوتید

نونوکلئوتیدها، مولکولهای آلی شامل نوکلئوزید و فسفات می باشند. آن ها به عنوان واحدهای مونومری، پلیمرهای نوکلئیک اسیدی: دئوکسی ریبونوکلئیک اسید (دی‌ان‌ای) و ریبونوکلئیک اسید (آران‌ای) را تشکیل می‌دهند، که هردو مولکول‌های زیستی اساسی در تمام اشکال حیات روی زمین می باشند. نوکلئوتیدها از طریق رژیم غذایی به دست آمده و هم‌چنین در کبد از طریق مواد غذایی رایج سنتز می‌گردند.[10]

نوکلئوتیدها از سه زیر واحد مولکولی تشکیل شده اند: یک باز نوکلئوتیدی، یک قند پنج کربنه پنتوز (ریبوز یا دئوکسی ریبوز)، و یک گروه فسفات شامل یک تا سه فسفات. چهار باز نوکلئوتیدی دی‌ان‌ای شامل: گوانین، آدنین، سیتوزین و تیمین می باشند؛ در آران‌ای، اوراسیل به جای تیمین استفاده می‌گردد.



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

2.1 آران‌ای

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

نوعی از آران‌ای اطلاعات را از دی‌ان‌ای به سیتوپلاسم حمل می‌کند؛ به این نوع آران‌ای که اطلاعات را از دی‌ان‌ای به ریبوزوم‌ها حمل

[1] Selective Breeding

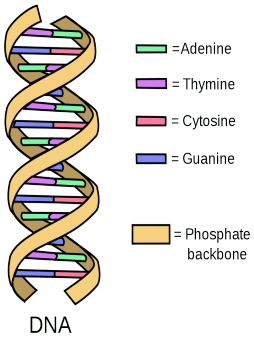
[2] RiboNucleic Acid

می‌کند، آران‌ای پیک یا پیامبر(mRNA) می‌گویند. نوعی دیگر از ،آران‌ای آران‌ای 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، آران‌ای ریبوزومی (rRNA) است که در ساختار ریبوزوم‌ها شرکت دارد؛ این موضوع به این معناست که ریبوزوم (رناتن) ها متشکل از پروتئین ها و آران‌ای های ریبوزومی هستند.

3.1 دی‌ان‌ای

دئوکسی ریبو نوکلئیک اسید[3] به اختصار دی‌اِن‌اِی یک مولکول متشکل از دو زنجیره پلی نوکلئوتیدی است که به دور یکدیگر می‌پیچند که دارای دستورالعمل‌های ژنتیکی است که 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 چهار درشت‌مولکول‌های اصلی و ضروری برای همه اشکال شناخته شده حیات هستند.

دو رشته دی‌ان‌ای 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 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 دی‌ان‌ای 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 دی‌ان‌ای، پیریمیدین ها تیمین و سیتوزین هستند. پورین‌ها آدنین و گوانین هستند.



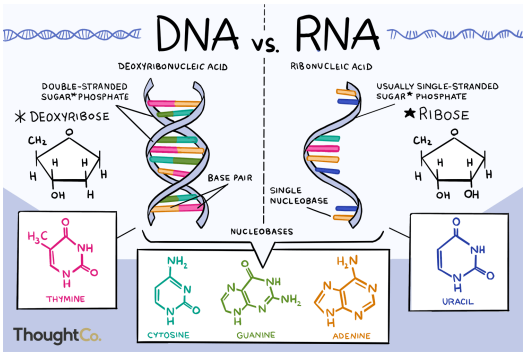
شکل 2.1: شکل دو بعدی دی‌ان‌ای [19]

هر دو رشته دی‌ان‌ای اطلاعات بیولوژیکی یکسانی را ذخیره می‌کنند. این اطلاعات زمانی که دو رشته از هم جدا می‌شوند، تکرار می شود. بخش بزرگی از دی‌ان‌ای (بیش از 98٪ برای انسان) بی‌کد[4] است، به این معنی که این بخش‌ها توالی‌های پروتئین را کد نمی‌کنند. دو رشته دی‌ان‌ای 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 آران‌ای با استفاده از رشته‌های دی‌ان‌ای به‌عنوان یک الگو در فرآیندی به نام رونویسی ایجاد می‌شوند که در آن بازهای دی‌ان‌ای با بازهای مربوطه خود مبادله می‌شوند، به جز در مورد تیمین ،(T) که آران‌ای جایگزین اوراسیل (U) می‌شود. تحت کد ژنتیکی، این رشته‌های آران‌ای توالی اسیدهای آمینه درون پروتئین‌ها را در فرآیندی به نام ترجمه مشخص می‌کنند.

1.3.1 تفاوت‌های دی‌ان‌ای و آران‌ای

تفاوت‌ها:

- دی‌ان‌ای برعکس آران‌ای از هستهٔ سلول خارج نمی‌شود.
- آران‌ای بدون ژن می‌باشد.
- دی‌ان‌ای در ذخیره و آران‌ای در انتقال اطلاعات وراثتی و در ساختار ریبوزوم نقش دارد.
- مولکول دی‌ان‌ای دو رشته‌ای در هم تنیده اما مولکول آران‌ای تک‌رشته‌ای است.



شکل 3.1: مقایسه دی‌ان‌ای و آران‌ای [22]

[3] Deoxyribonucleic acid
[4] non-coding

- در دی‌ان‌ای باز آلی یوراسیل و در آران‌ای باز آلی تیمین شرکت ندارد (U در دی‌ان‌ای و T در آران‌ای).
- قند پنج‌کربنه موجود در دی‌ان‌ای را دئوکسی ریبوز و در آران‌ای 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.

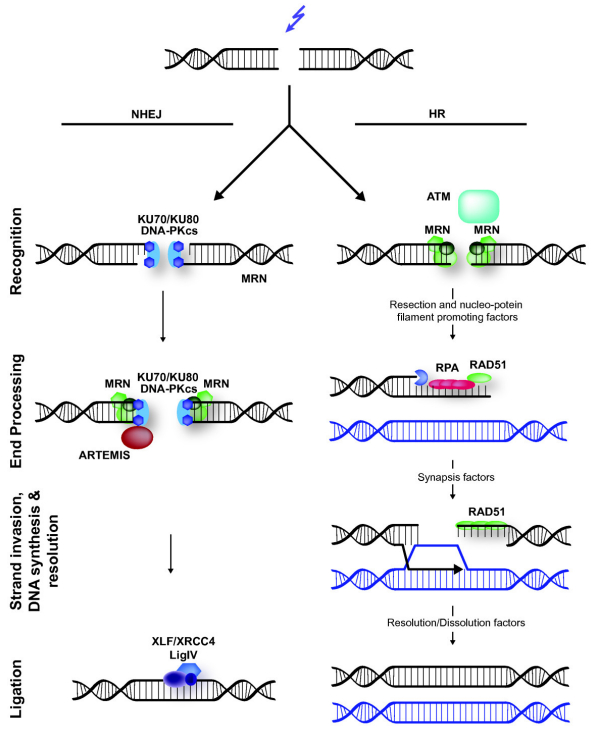
شباهت‌ها:

- 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.

4.1 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 رادیواکتیو بر روی گیاهان به تغییر تصادفی بر روی ژنوم دست یافتند. 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 پالیندروم فاصله‌دارِ منظمِ خوشه‌ای[7] (CRISPR). کلید ویرایشِ ژنوم ایجادِ شکستِ دو رشته‌ی دی‌ان‌ای در نقطه مورد نظر است و این سه روش مبتنی بر شکست درست دی‌ان‌ای در نقطه مورد نظر مهندسی شدند.



شکل 4.1: مکانیزم ترمیم دی‌ان‌ای، عکس گرفته شده از ویکی‌پدیا.

1.4.1 شکست و تعمیر دی‌ان‌ای

شکل رایج ویرایش ژنوم بر مفهوم مکانیک ترمیم شکست دو رشته‌ای دی‌ان‌ای[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 دی‌ان‌ای استفاده می‌کند، در حالی که HDR دقیق‌تر از یک توالی همولوگ به عنوان الگویی برای بازسازی توالی‌های دی‌ان‌ای 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 دی‌ان‌ای 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

[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

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 دی‌ان‌ای 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 دی‌ان‌ای می شود. ایندل ها اغلب منجر به تغییر اساسی در دی‌ان‌ای می شوند، به‌طوری که دی‌ان‌ای 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 DBS، دی‌ان‌ای به این روش تعمیر نشود.

HDR

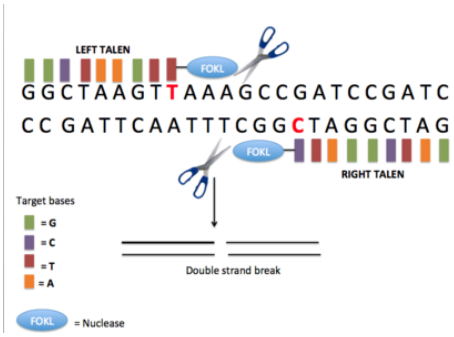
تعمیر هدایت شده همولوژی (HDR) مکانیزمی در سلول ها برای ترمیم ضایعات دی‌ان‌ای دو رشته‌ای از یک نسخه مشابه دی‌ان‌ای برای ترمیم استفاده می کند. رایج ترین شکل HDR نوترکیبی همولوگ است. مکانیسم HDR تنها زمانی می تواند توسط سلول استفاده شود که یک قطعه همولوگ از دی‌ان‌ای در هسته وجود داشته باشد، عمدتاً در فاز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

Zinc finger nucleases (ZFN) 2.4.1

نوکلئاز انگشت روی یا ZFN اولین سیستم پروتئینی متصل شونده به دی‌ان‌ای 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 دی‌ان‌ای 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 دی‌ان‌ای 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 دی‌ان‌ای باید دایمریزه شده باشد، بنابراین یک جفت از ZFN برای هدف گیری و برش دی‌ان‌ای 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 دی‌ان‌ای 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.

TALEN 3.4.1

نوکلئازهای رونویس مؤثر-مانند فعال‌کننده [11] (TALENs) پپروتئین های متصل شونده به دی‌ان‌ای با آرایه تکراری 33 یا 34 اسید آمینه هستند. TALEN ها آنزیم های محدودکننده مصنوعی هستند که از ادغام حوزه برش دی‌ان‌ای یک نوکلئاز با دامنه های TALE طراحی شده اند، که می توانند به طور خاص یک توالی دی‌ان‌ای منحصر به فرد را شناسایی کنند. این پروتئین‌های ادغام شده به‌عنوان قیچی دی‌ان‌ای 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.[53] این تکنولوژی را می‌توان برای تغییر هر نقطه از دی‌ان‌ای استفاده کرد. TALEهای 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، به استثنای دی باقیمانده متغیر تکرار شونده [12] (RVDs) در موقعیت های اسید آمینه 12 و 13، هر تکرار حفظ می شود. RVD ها توالی دی‌ان‌ای را تعیین می کنند که TALE به آن متصل می شود. این تناظر ساده یک به یک بین تکرارهای TALE و توالی دی‌ان‌ای مربوطه باعث می‌شود روند مونتاژ آرایه‌های تکراری برای تشخیص توالی‌های دی‌ان‌ای جدید ساده باشد. این TALEها را می توان با کاتالیزوری از یک نوکلئاز از دی‌ان‌ای به نام FokI ، ادغام کرد تا با آن هاTALEN را ساخت. ساختارهای TALEN توالی‌های دی‌ان‌ای را فقط در



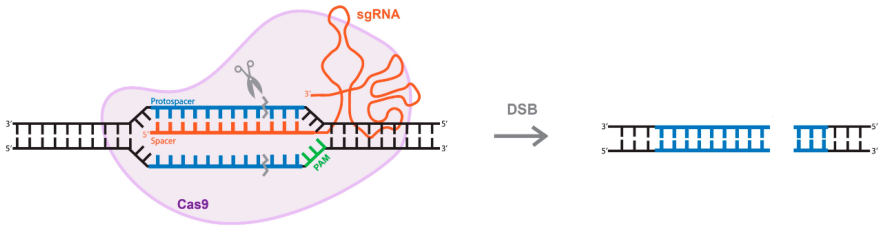
شکل 5.1: مکانیزم TALEN [20]

[11] Transcription activator-like effector nucleases
[12] Repeat Variable Di-residues

مکان‌های از پیش انتخاب شده متصل می‌کنند و می‌شکنند. هدف TALEN را می توان بر اساس یک کد آسان پیش بینی کرد. با توجه به این که محل اتصال بیش از ۳۰ جفت نوکلئوتید است، نوکلئازهای TAL مختص هدفی یکتا هستند. هر نوکلئوتید منفرد در ژنوم در صورتی که در محدوده 6 جفت نوکلئوتید باشد، TALEN می‌تواند آن را ویرایش کند. 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 دی‌ان‌ای بالاتر است، (2) اثرات خارج از هدف کمتر است و (3) طراحی آن آسان تر است.[54]

5.1 کریسپر

Clustered Regularly Interspaced Short Palindromic Repeats یا به اختصار کریسپر (به انگلیسی: CRISPR) به معنی "تناوب‌هایِ کوتاهِ پالیندرومِ فاصله‌دارِ منظمِ خوشه‌ای" 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 دی‌ان‌ای 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 دی‌ان‌ای را در محل خاصی از ژنوم برش دهد.[23]



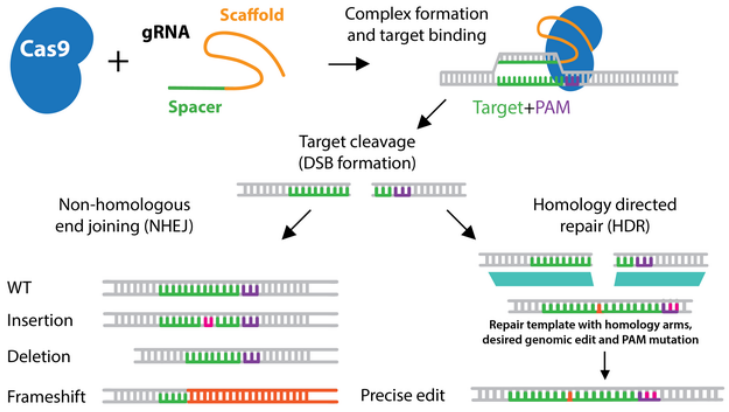
شکل 6.1: مکانیزم ساده شده‌ای از CRISPR [18]

1.5.1 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 آران‌ای کوچک شناساگر توالی خاص تکیه می‌کنند و اسیدهای نوکلئیک خارجی را خاموش می‌کنند. [28] Francisco Mojica و همکارانش در سال ۱۹۹۳ تکرارهای مشابهی را در چندین گونه میکروبی دیگر یافتند.[39] 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 آران‌ای 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 پرادازش 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 آران‌ای کریسپر فعال‌کننده ترانس است (tracrRNA) که با تکرارهای مستقیم هیبرید می‌شود و یک آران‌ای دوپلکس را تشکیل می‌دهد و توسط RNase III درونی و نوکلئازهای ناشناخته دیگر شکسته و پردازش می‌شود. CrRNA های 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 دی‌ان‌ای 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 دی‌ان‌ای و RNA را می‌شکنند. درمقابل، سیستم نوع II فقط نیاز به Cas9 برای تخریب دی‌ان‌ای جفت شده با آران‌ای راهنما دوپلکس خود دارد که این آران‌ای راهنما حاوی ترکیبی از CrRNA-tracrRNA است.[۵]

2.5.1 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 آران‌ای و cas تقسیم کرد که cas در آن وظیفه جدا کردن دو رشته دی‌ان‌ای را از هم دارد و آران‌ای که هدف را مشخص و قیچی می‌کند. برای این که دقیقا نقطه شکست دی‌ان‌ای مشخص شود cas نیاز به یک علامت 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 دی‌ان‌ای، دو مکانیزم برای تعمیر آن وجود دارد. دانشمندان تکنولوژی‌های CRISPR مختلفی را برای افزایش احتمال تعمیر HDR ایجاد کرده‌اند که هر یک ویژگی‌های خاص خود را دارند ولی ما در پژوهش خود ساده‌ترین مورد آن یعنی cas9 به همراه یک آران‌ای 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 [8]

3.5.1 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 دی‌ان‌ای هدف خواهد داشت و هیچ همسانی در جای دیگری در ژنوم وجود ندارد یعنی دقیقا 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 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 دی‌ان‌ای نیک تولید می کند.

بنابراین، دو نیکاز که رشته‌های دی‌ان‌ای مخالف را هدف قرار می دهند برای تولید DSB در دی‌ان‌ای 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 دی‌ان‌ای 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 دی‌ان‌ای 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.

4.5.1 تاثیرگذاری

تاثیرگذاری در یک طرح CRISPR احتمال شکست دی‌ان‌ای 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 دی‌ان‌ای 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 دی‌ان‌ای گلیکوزیلاز (UGI) یا Gam مشتق از باکتریوفاژ (Mu-GAM) انجام شده است.

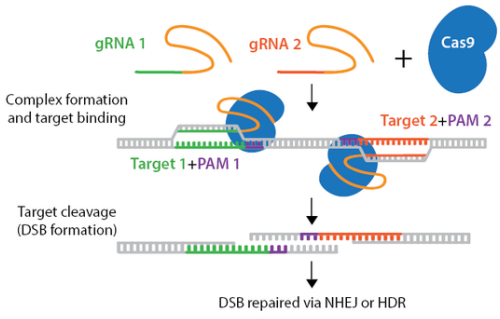
5.5.1 انواع کریسپر

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 دی‌ان‌ای ژنوم را تجزیه می‌کند. نوع II CRISPR/Cas اولین سیستم برای مهندسی ژنوم، با نوع V در ۲۰۱۵ بود.

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 دی‌ان‌ای خارجی 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 [8]

نیکاز CRISPR/Cas جهش‌یافته، به‌جای شکستگی‌های دو رشته‌ای ایجاد شده توسط آنزیم‌های Cas، شکستگی‌های تک‌رشته‌ای با هدف gRNA را در دی‌ان‌ای ایجاد می‌کنند. برای استفاده از جهش نیکاز، به دو gRNA نیاز دارید که رشته‌های مخالف دی‌ان‌ای 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.

مراجع این فصل: [4–9, 11–17, 23, 25, 32, 33, 37, 37, 38, 43, 46, 51]

در این پژوهش ما به حل مسئله تأثیرگذاری می‌پردازیم، و در ادامه روش‌هایی که برای حل مسئله استفاده کرده‌ایم را برای شما بازگو می‌کنیم. ابتدا کارهای پیشین را توضیح می‌دهیم و سپس مشکلات آن را توضیح می دهیم. در ادامه روش‌هایی که برای حل مسئله استفاده کرده‌ایم را که چه به شکست و چه موفق بوده است را توضیح می‌دهیم.

جدول 1.1: خلاصه‌ای از اصطلاحات به کار برده و تعریف آنها

اصطلاح	تعریف
ویرایشگر پایه (Base editor)	ادغام یک پروتئین Cas به یک دآمیناز که تبدیل مستقیم باز در آران‌ای یا دی‌ان‌ای را بدون شکست دو رشته دی‌ان‌ای 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 دفاعِ 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 دی‌ان‌ای
dCas9	Nuclease dead Cas9, شکل آنزیمی غیر فعال .Cas9 می تواند متصل شود، اما نمی تواند دی‌ان‌ای 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 دی‌ان‌ای 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 دی‌ان‌ای هدف استفاده می‌شود، نمی‌شود.
gRNA targeting sequence	۲۰ نوکلئوتید قبل از توالی PAM در دی‌ان‌ای 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 دی‌ان‌ای که از یک الگو برای ترمیم نیک های دی‌ان‌ای 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 دی‌ان‌ای 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 دی‌ان‌ای هدف ضروری است
PCR	Polymerase Chain Reaction; برای تقویت و خوانا شدن یک توالی خاص از دی‌ان‌ای 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.

جدول 2.1: برخی از انواع کریسپر و PAM آن

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

R = G or A, Y = C or T, W = A or T, N = A or C or G or T

فصل 2

کارهای پیشین

مطالعات زیاد و متعددی روی مشکلات کریسپر انجام شده است ولی در اینجا ما آن‌ها را به دو دسته مختلف تقسیم می کنیم. روش‌های مستقیم که در آن‌ها دانشمندان به رابطه‌های مستقیم بین مکانیزم‌ها مختلف و تاثیر آنها روی دقت و حساسیت طرح‌ها مورد بررسی قرار داده‌اند. و دسته دوم روش‌های یادگیری ژرف می‌باشند که برای پیش‌بینی تاثیر و حساسیت طرح‌ها مورد استفاده قرار گرفته اند.

1.2 روش‌های مستقیم

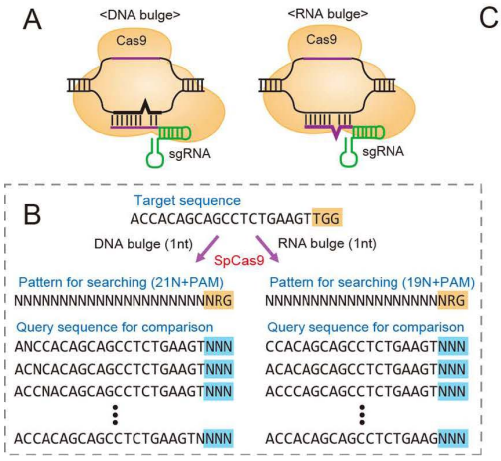
1.1.2 Chopchop [2, 3, 34]

این مقاله که الگوریتم خود را سه بار بروزرسانی کرده است، به عنوان ورودی رشته دی‌ان‌ای ورودی و یا اسم ژن یا مختصات 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 دسترس بودن هدف در احتمال شکسته شدن دی‌ان‌ای 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 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 دی‌ان‌ای، 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 دی‌ان‌ای 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 [31, 48]

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 انتخابی خود برای RNA مشتق شده ازCRISPR/Cas نوع II است، که در حال حاضر به طور گسترده برای تحقیقات زیست پزشکی و بیوتکنولوژی استفاده می‌شود. Cas-Designer به سرعت فهرستی از تمام توالی های آران‌ای راهنمای ممکن در یک توالی دی‌ان‌ای ورودی داده شده ارائه می دهد و off-target آنها را در ژنوم انتخابی مشخص می‌کند. علاوه بر این، برنامه امتیاز خارج از چارچوب را به هر نقطه از هدف اختصاص می دهد تا به کاربران کمک کند مناطق مناسب برای Knockout ژن انتخاب کنند. Cas-Designer نتایج را در یک جدول تعاملی نشان 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 StCas9 (Cong et al., 2013), 50-NNNNGMTT-30 برای NmCas9 (Hou et al., 2013) و 50-NNGRRT-30 برای SaCas9 (Ran et al., 2015)] در یک توالی دی‌ان‌ای 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: ACCATAgTgAAAACTCCACAAGGG	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-TCCAaAATGG	chr8	128284008	-	2	1	Info at Ensembl
RNA	crRNA: ACCTATTTAAAACTCCACAANGG DNA: ACCTATTgAAAAtTCC-CAATGG	chr8	103798023	+	2	1	Info at Ensembl

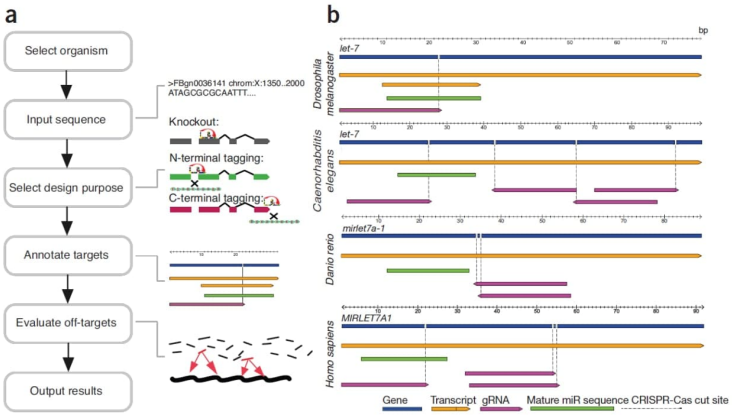
شکل 1.2: (الف) شماتیک مکان‌های off-tagets را با برآمدگی دی‌ان‌ای یا آران‌ای نشان می‌دهد. (ب) استراتژی برای برآمدگی 1-nt DNA یا آران‌ای بر اساس .Cas-OFFinder ج) یک مثال از یک جدول خروجی Cas-Designer تمام gRNA های ممکن را از توالی های ورودی به همراه اطلاعات مفید (بالا) نشان می دهد. اگر کاربر روی رنگ آبی کلیک کند عدد، کلمه یا عبارت، اطلاعات دقیق تری مانند اهداف برآمدگی دی‌ان‌ای (وسط) یا آران‌ای (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" [48]

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 با ++C نوشته است. 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 [24]

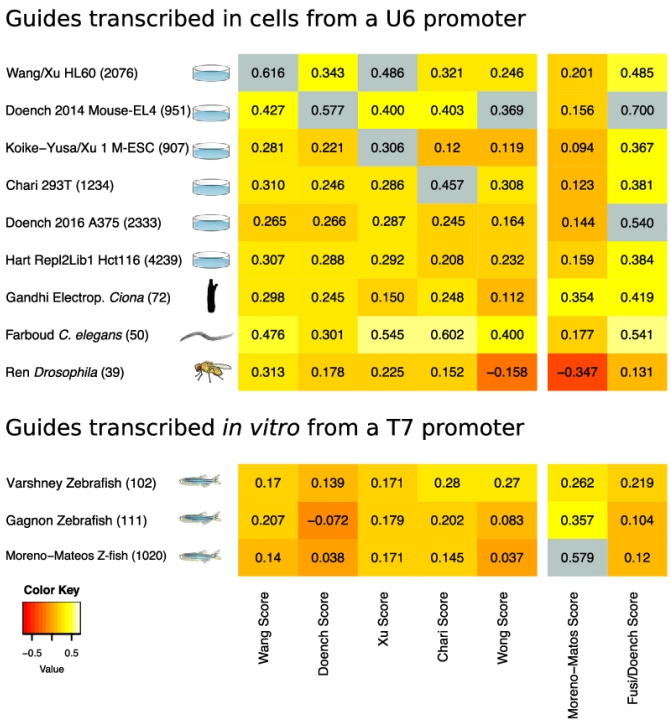
در اینجا ما E-CRISP، یک برنامه وب برای طراحی توالی های gRNA را توصیف می کنیم. (الف) مراحل E-CRISP. ابتدا کاربر ارگانیسم و دنباله هدف را انتخاب می کند. این هدف می تواند یک نماد ژن، یک شناسه ENSEMBL یا یک توالی FASTA باشد. دوم، کاربر هدف آزمایش ویرایش را مشخص می کند. بسته به هدف، E-CRISP مناطق مختلفی از 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) آران‌ای 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 دی‌ان‌ای 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 [24]

4.1.2 CRISPOR [30]

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 دی‌ان‌ای 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، آران‌ای 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 ۰.۱ ٪ 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 دست 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 می توان با زمان مناسب پیش‌بینی مناسبی ارائه داد.

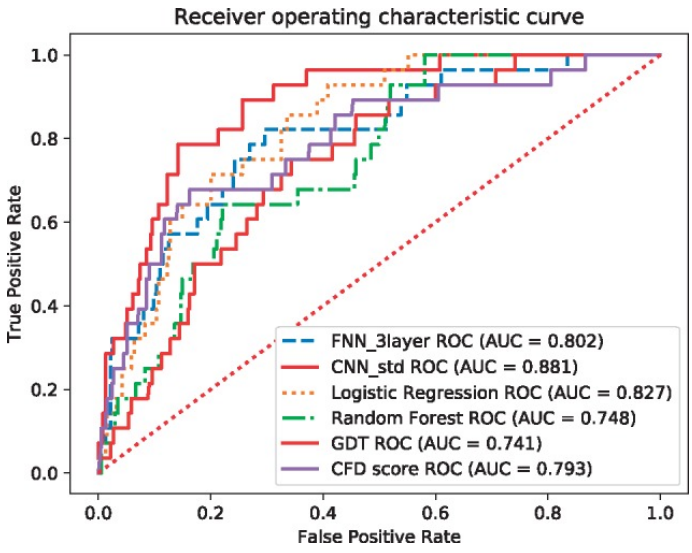


شکل 3.2: هم بستگی اسپیرمن بین امتیاز موثر بودن و داده‌ها [30]

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 پُر پژوهش‌ای 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، این پژوهش 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 CROP-IT، MIT، CFD 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



شکل 4.2: هم بستگی اسپیرمن بین امتیاز موثر بودن و داده‌ها [52]

رقابتی الگوریتم های پیشنهادی است. تحلیل های اضافی برای بررسی دلایل زمینه ای از دیدگاه های مختلف انجام می شود.

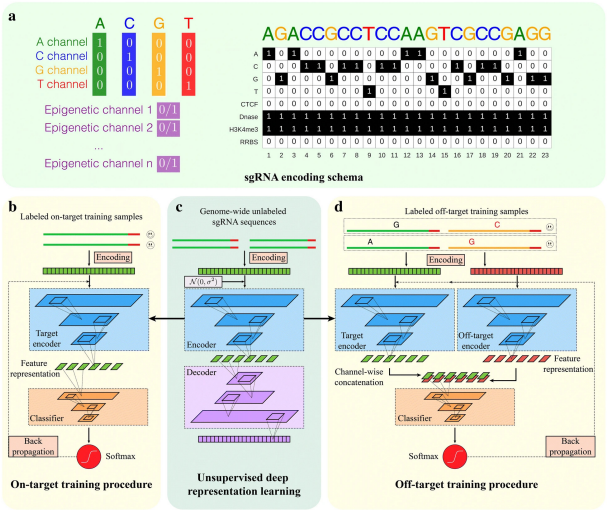
2.2.2 CCTop [49]

این روش برای اینکه طرح‌های مختلف که به صورت N20NGG هستند را دسته بندی می‌کند، ابتدا با آزمایش‌های عملی طرح‌ها را به دو کلاس موثر و ناموثر دسته‌بندی کرده‌اند. آزمایش به این گونه بود که در محیط آزمایشگاهی طرح را به ژن تزریق می کردند و برای هر طرح را با تعداد هدف‌های تغییر کرده در طول زمان یادداشت کرده اند. 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 تعیین کرده است. این طرح جایگاه هر نیکلوتید 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 Yها امتیاز آن‌ها باشد داریم:
پس از تمرین این مدل، برای بهتر فهمیدن مدل روی دیتای آزمایشگاهی و مدل خود آزمایش‌های آماری از جمله ارتباط 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:

- قرار گرفتن نیکلوتید G در موقعیت های ۱- و ۲- نسبت به PAM در CAS9 باعث افزایش تاثیرگذاری می‌شود.
- قرار گرفتن نیکلوتید T در چهار موقعیت نزدیک به PAM باعث کاهش تاثیرگذاری می‌شود.
- نوکلئوتیدهای رشته ′۵ به ′۳ تاثیرگذار هستند، در حالی که رشته مکمل تاثیر قابل توجهی ندارد.
- قرار گرفتن نیکلوتید C در موقعیت ۳- در CAS9 باعث افزایش تاثیرگذاری می‌شود.
- قرار گرفتن نیکلوتید A در موقعیت -5 تا -12 باعث افزایش تاثیرگذاری می‌شود.
- قرار گرفتن نیکلوتید G در موقعیت های -14 تا -17 باعث افزایش تاثیرگذاری می‌شود.

3.2 DeepCRISPR

DeepCRISPR ، یک پلتفرم محاسباتی جامع برای یکپارچه سازی پیش‌بینی ناحیه sgRNA روی هدف و خارج از هدف در یک چارچوب با یادگیری عمیق، با استفاده از پیشرفته‌ترین ابزارهای موجود در سیلیکون است. DeepCrispr [26] علاوه بر ویژگی های توالی دی‌ان‌ای، چهار ویژگی اپی ژنتیکی را معرفی کرد و به طور خودکار اطلاعات معتبر را با استفاده از اصل Auto-encoder استخراج می کند. 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 بی تاثیر نیست.



شکل 5.2: DeepCRISPR [26]

با توجه به این که کارهای پیشین روی اورگان‌های مختلف آموزش داده شده اند در اورگان‌هایی که تا به حال ندیده اند، نتایج خوبی ندارند و همین‌طور با اینکه دقت این مدل‌ها بالا است، هنوز به دقتی قابل اعتماد تبدیل نشده‌اند، در نتیجه در این پژوهش ما سعی می‌کنیم که برای مشکلات راه حلی بهتر ارائه دهیم.

جدول 1.2: خلاصه‌ای از کارهای پیشین

Method	Input	Enzyme	Organism	On-Target Scoring Method	Off-Target Scoring Method	Features
CHOPCHOP	GeneID Coordinates Sequence	SpCas9; SpCas9n; Cas12a (Cpf1); CasX; Cas13 (C2C2); TALEN	Variety	Doench et al. 2014; Doench et al. 2016; Chari et al. 2015; Xu et al. 2015; Moreno-Mateos et al. 2015; G20	MIT specificity score; Cong et al., 2013	Designs primers for the edited site amplification; restriction sites map; exon-intron map; Integrates Shen et al. 2018 predictions of repair profile
CRISPOR	Coordinates Sequence	SpCas9; SpCas9-HF1; eSpCas9 1.1; ScCas9; iSpyMacCas9; SaCas9; xCas9; SaCas9-KKH; SpCas9-VQR; NmeCas9; SpCas9-VRER; StCas9; CjCas9; AsCas12a (Cpf1); LbCas12a (Cpf1)	Variety	Doench et al. 2016 Chari et al. 2015; Xu et al. 2015; Wu-Crisp Doench et al. 2014; Wang et al. 2014 Moreno-Mateos et al. 2015; Azimuth in-vitro crisprRank	MIT Specificity Score; CFD Specificity score	Designs primers for the edited site amplification; restriction sites map; provides sequences for in vitro expression or cloning of designed sgRNAs; Integrates Bae et al. 2014 predictions of repair profile and Chen et al. 2018 frameshift prediction
E-CRISP	GeneID Sequence	SpCas9	Variety	Heighwer et al. 2014; Doench et al. 2014; Xu et al. 2015	Bowtie2	Includes genetic variation
CasFinder CasDesigner	Coordinate Sequence	SpCas9; StCas9; NmeCas9	Homo sapiens Mus musculus	Aach et al. 2014	Exome-wide catalog of Cas9 cleavage sites	Features
CCTop	Sequence	SpCas9; SpCas9-VQR; SpCas9-VRER; AsCas12a (Cpf1); LbCas12a (Cpf1); FnCas12a (Cpf1); SaCas9; StCas9; NmeCas9; TdCas9	Variety	CRISPRater	Stemmer et al. 2017	Includes genetic variation
DeepCRISPR	Sequence	SpCas9	Homo sapiens	Chuai et al. 2018	Chuai et al. 2018	Integrates the epigenetic information in different cell types

فصل 3

روش‌های پیشنهادی

ابتدا یک بار دیگه مسئله را مدل سازی می‌کنیم، فرض کنید یک رشته ۲۳ تایی از نوکلوتیدها را در اختیار داریم، پس اگر N یک نوکلوتید دلخواه باشد، رشته دلخواه به صورت زیر است:

$$NNNNNNNNNNNNNNNNNNNNNNN$$

از آنجایی که در پژوهش ما فقط از cas9 استفاده می‌شود دو نوکلوتید آخر باید G باشد پس داریم:

$$NNNNNNNNNNNNNNNNNNNNNGG$$

به این رشته ۲۳تایی ترکیب ۲۰تایی sgRNA و ۳تایی PAM می گویند که برای cas9 ، حتما باید NGG باشد. تمام کارهای پیشین به روشی این رشته ۲۳تایی را با توجه به ویژگی‌های مختلف یا اینکدینگ به متغیر کمی تبدیل کرده‌اند. پس به عنوان ورودی این رشته و ویژگی‌های مختلف را استفاده می‌کنند تا به عنوان خروجی یک عدد بین صفر و یک به عنوان امتیاز تاثیرگذاری به ما می دهد. داده‌های امتیاز تاثیرگذاری به صورت آزمایشگاهی توسط پژوهشگران با ادغام طرح کریسپر با سلول هدف و یادداشت نتایج آن در طول زمان بدست می‌آید. این نتایج با توجه به پژوهش متفاوت می باشد که یک مشکل برای ادغام داده‌ها می باشد از جلمه دو مورد از نتایج نظارت شده هنگام این آزمایش‌ها، شمردن ایندل‌های و شمردن تعداد شکست‌های دی‌ان‌ای است. همچنین شایان ذکر است که تاثیرگذاری sgRNA ها در جانورهای مختلف و اورگان‌های مختلف متفاوت است ولی با توجه به پژوهش‌های صورت گرفته مانند DeepCRISPR [26] هم چنان با نادیده گرفتن این اطلاعات و تمرکز روی رشته sgRNA نیز می‌توان پیش‌بینی‌های مفیدی انجام داد. با در نظر گرفتن این فرضیات نیز هنوز فضای مسئله فضای بسیار بزرگی است از آنجا که تعداد طرح‌های ممکنه برای با تعداد جایگشت با تکرار ۴ شی در ۲۱ خانه یا همان $4^{21} \approx 4{,}3980465 \times 10^{12}$ است. در نتیجه با دیدن حدود ۱۰ هزار یا حتی ۱۰۰ هزار نمونه نمی‌توان نتیجه‌ی عمومی درباره‌ی این مسئله گرفت. برای درست کردن روشی عمومی به داده‌های زیاد و دقیق نیاز است که در حال حاضر در دسترس نیستند و همین‌طور معمولا این دادگان برای یک روش خاص تهیه شده‌اند که یعنی بعضی از ویژگی‌های مورد نیاز برای بعضی داده‌ها موجود و برای برخی دیگر موجود نیستند، از آنجا که قادر به درست کردن مجموعه داده مناسب و عمومی نبوده‌ایم، سعی کردیم با کمترین ویژگی‌ها مدلی بسازیم که بهترین دقت را داشته باشد، یعنی فقط دنباله دی‌ان‌ای و نام اروگان. در این پژوهش، ما دو ایده برای تبدیل متغیر کیفی sgRNA به متغیر کمی داشتیم. ایده اول استفاده از نتایج کارهای پیشین به عنوان نمایش بردار کمی sgRNA بود و ایده دوم استفاده از مدل‌های attention و transformer ها برای بدست آوردن یک اینکدینگ مناسب است.

از آنجا که بیشتر کارهای پیشین ویژگی‌های دیگر مورد نیاز خود را از ورودی متدها نمی‌گرفتند یک روش مناسب برای حذف این ویژگی‌های اضافه و کمک به عمومی شدن مدل، ادغام متدهای مختلف است ولی با توجه به نتایج ادغام این مدل‌ها به تنهایی کافی نیست، در نتیجه برای رفع این مشکل، ما از روشی نوین که ایده‌ای مشابه به Stacked Generalization [55] دارد استفاده می‌کنیم تا با مجموعه دادگان کم دقت بهتری بدست‌آوریم. می‌توان به روش بدست آمده مانند اصلاح اشتباهات یک مدل توسط مدل دیگر نگاه کرد که در آن از چند مدل مختلف چندین نمونه داریم که همگی باهم ادغام می‌شوند تا بهتر نتیجه از یک مدل بدست بیاید و برای بدست آمدن بهترین نتیجه هر مدل برای این که متر مناسب و معلومی وجود ندارد چندین یک از پر کاربردترین خطاها را استفاده کردیم و مدل را با آن فاینتون کردیم و با رایگیری بین همه خطاهای مختلف نتیجه‌ی پیشبینی مدل را به عنوان بهترین جواب مدل در نظر گرفتیم. با بدست آمدن بهترین نمونه از هر مدل، مدل‌ها را با هم ادغام می کنیم تا با اصلاح یک دیگر بهترین دقت را به ما ارائه دهند. واضح است که ممکن است دو مدل مختلف نقاط ضعف و قوت مشترکی داشته باشند که در این صورت روش ذکر شده مفید نخواهد بود.

1.3 Learning 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. [40, 42, 47] 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. [27]

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 بخش نتایج، نمونه‌هایی از این روش‌ها را نشان می‌دهیم.

در مرحله بعدی با جنگل‌های تصادفی سعی 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 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، رگرسیون با جنگل تصادفی، درختان بسیار تصادفی، حداقل مربعات معمولی، تقویت گرادیان 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-Designer ، 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 standard golden برای این داده‌ها پیدا کنیم، که متاسفانه قادر به این کار نشدیم.

2.1.3 رگرسیون با جنگل تصادفی

رگرسیون با جنگل تصادفی [57] یک الگوریتم یادگیری نظارت شده است که از روش یادگیری ادغامی برای رگرسیون استفاده می کند.

مقدمات: آموزش درخت تصمیم

درخت تصمیم روش مشهوری برای انواع مختلفی از وظایف یادگیری ماشین به حساب می آید. با این حال در بسیاری موارد دقیق نیستند.

در کل، معمولا درخت تصمیمی که بیش از حد عمیق باشد الگوی دقیق نخواهد داشت: دچار بیش برارزش شده، و دارای سوگیری پایین و واریانس بالا می‌باشد. جنگل تصادفی روشی است برای میانگین گیری با هدف کاهش واریانس با استفاده از درخت‌های تصمیم

عمیقی که از قسمت‌های مختلف داده آموزشی ایجاد شده باشند. در این روش معمولا افزایش جزئی سوگیری و از دست رفتن کمی از قابلیت تفسیر اتفاق افتاده اما در کل عملکرد مدل را بسیار افزایش خواهد داد.

کیسه‌گذاری درختان

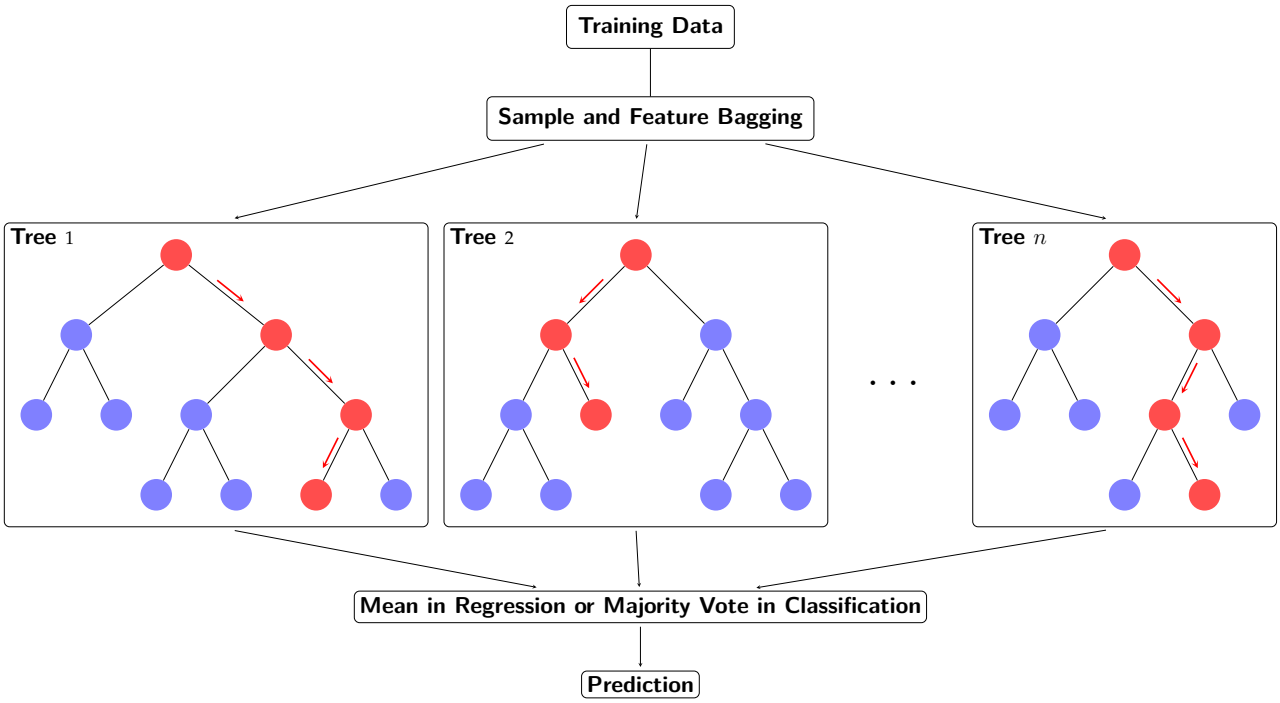
مجموعه داده را با D نمایش میدهیم، $D = (x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ و B درخت تصادفی با ایجاد B داده جدید از D ایجاد می‌کنیم. مدل نهایی با میانگین گرفتن یا رأی‌گیری بین درختان کار می‌کند. جزئیات این الگوریتم ذیلاً آمده است:

برای B تا $b = 1$:

- n نمونه با جایگزینی از داده D انتخاب می‌کنیم و این نمونه‌ها را در مجموعه داده D_b قرار می‌دهیم. از آنجا که نمونه‌گیری با جایگزینی صورت می‌گیرد یک نمونه ممکن است چندین بار انتخاب شود.
- یک درخت تصادفی به اسم T_b با D_b به روش پایین می‌سازیم:
هر دفعه برای پیدا کردن بهترین متغیر ابتدا یک تعداد مشخصی از متغیرها را کاملا به صورت تصادفی انتخاب می‌کنیم (مثلا m متغیر اول به مسئله داده شده‌است، و معمولاً با جذر تعداد متغیرها برابر است) و از میان آن‌ها بهترین متغیر را انتخاب می‌کنیم.

در مسئله رگرسیون مدل نهائی، میانگین تمامی درخت‌ها است یعنی $F(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$. از طرفی دیگر در مسئله دسته‌بندی با رأی‌گیری بین درختان به جواب نهائی می‌رسیم.

این نوع ترکیب مدل‌ها جواب بهتری به ما می‌دهد زیرا گوناگونی و تنوع مدل‌ها را افزایش می‌دهد بدون این که بایاس را افزایش دهد. این بدین معناست که زمانی که پیش‌بینی تکی از یک درخت دارای نویز بالایی درون مجموعه دسته آموزش دیده‌اش باشد، در میانگین بسیاری از درخت‌ها این نویز وجود نخواهد داشت. به شکل ساده آموزش درختان به صورت تکی می‌تواند درخت‌های در ارتباط قوی تری را ارائه دهد. بوت استرپ کردن نمونه، روشی برای یکپارچه‌تر کردن درخت‌ها با نمایش مجموعه داده‌های آموزش دیده گوناگون است.



3.1.3 درختان بسیار تصادفی

در درختان بسیار تصادفی[58]، یک قدم تصادفی بیشتر دارد. همانند جنگل‌های تصادفی، زیرمجموعه‌ای تصادفی از متغیرها کاندید می‌شود، اما به جای جستجوی بهترین آستانه، آستانه‌ها به‌طور تصادفی برای هر متغیر کاندید شده ترسیم می‌شود و بهترین این آستانه‌های تصادفی تولید شده به عنوان آستانه تقسیم انتخاب می‌شوند. این امر عموما به کاهش کمی بیشتر واریانس مدل منجر می‌شود و باعث افزایش کوچکی در بایاس می‌شود.

4.1.3 حداقل مربعات معمولی

در آمار، حداقل مربعات معمولی (به انگلیسی: Ordinary Least Squares) (به اختصار OLS)، روشی است برای برآورد پارامترهای مجهول در مدل رگرسیون خطی از طریق کمینه کردن اختلاف بین متغیرهای جواب مشاهده شده در مجموعه داده است. فرض کنید که n مشاهده‌ی $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$ داریم. هر مشاهده i شامل یک پاسخ اسکالر y_i و یک بردار ستونی $\mathbf{x}_i$ از پارامترهای p (رگرسور) ، به عنوان مثال، $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{ip}]^{\mathsf{T}}$. در یک مدل رگرسیون خطی، متغیر پاسخ، y_i، یک تابع خطی از رگرسورها است:

$$y_i = \beta_1\, x_{i1} + \beta_2\, x_{i2} + \cdots + \beta_p\, x_{ip} + \varepsilon_i,$$

که از به عنوان بردار به آن نگاه کنیم داریم:

$$y_i = \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_i,$$

به طوری که $\mathbf{x}_i$ بردار ستونی از i-امین مشاهده همه متغیرهای است و $\boldsymbol{\beta}$ یک بردار $p \times 1$ از پارامترهای ناشناخته است. و اسکالار ε_i نشان دهنده متغیرهای تصادفی مشاهده نشده (خطاهای) مشاهده i-ام است. ε_i تأثیرات توضیح‌دهنده‌های y_i توسط $\mathbf{x}_i$ نشان می‌دهد. این مدل را می توان به صورت نماد ماتریسی نیز نوشت:

$$\mathbf{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

به طوری که $\mathbf{y}$ و $\boldsymbol{\varepsilon}$ بردارهای $n \times 1$ هستند متغیرهای پاسخ و خطاهای n مشاهدات و X یک ماتریس $n \times p$ از رگرسیون‌ها است. گاهی اوقات ماتریس طراحی نیز نامیده می شود که سطر i-ام آن $\mathbf{x}_i^{\mathsf{T}}$ است و حاوی مشاهدات i-ام روی همه متغیرهای توضیحی است.

رگرسورها لازم نیست مستقل باشند: هر رابطه دلخواه بین رگرسیون ها می تواند وجود داشته باشد (تا زمانی که یک رابطه خطی نباشد). برای مثال، ممکن است مشکوک باشیم که پاسخ به صورت خطی هم به مقدار و هم به مربع آن بستگی دارد. در این صورت یک رگرسیون را که مقدار آن فقط مجذور رگرسیور دیگر است را در نظر می گیریم. در آن صورت، مدل در رگرسیور دوم درجه دوم خواهد بود، اما با این حال، همچنان یک مدل خطی در نظر گرفته می‌شود، زیرا مدل همچنان در پارامترهای خطی است.

از آنجایی که ε_i قابل محاسبه نیست برای استفاده از این روش معادله زیر را درنظر بگیرید:

$$\sum_{j=1}^{p} X_{ij}\beta_j = y_i,\ (i = 1, 2, \dots, n),\ n > p$$

چنین دستگاهی معمولاً راه جواب دقیق ندارد، بنابراین هدف در عوض یافتن ضرایب $\boldsymbol{\beta}$ است که نزدیکترین حالت به جواب باشد، به معنای دیگر حل مسئله کمینه سازی درجه دوم $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta})$, که در آن S برابر است با:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{p} X_{ij}\beta_j \right|^2 = \left\|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\right\|^2.$$

که اگر p ستون مستقل خطی باشند در این صورت دارای جواب یکتای:

$$(\mathrm{X}^{\mathsf{T}}\mathrm{X})\hat{\boldsymbol{\beta}} = \mathrm{X}^{\mathsf{T}}\mathbf{y}\,.$$

به عبارت دیگر:

$$\hat{\boldsymbol{\beta}} = \left(\mathrm{X}^{\mathsf{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathsf{T}}\mathbf{y}.$$

یا

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\boldsymbol{\varepsilon}.$$

5.1.3 تقویت گرادیان

در بسیاری از مسائل یادگیری تحت نظارت، یک متغیر خروجی y و یک بردار از متغیرهای ورودی x وجود دارد که با مقداری توزیع احتمالی به یکدیگر مرتبط هستند. هدف یافتن تابعی از $\hat{F}(x)$ است که به بهترین وجه متغیر خروجی را از مقادیر متغیرهای ورودی

تقریب می‌کند. این امر با معرفی تابع ضرر $L(y, F(x))$ و به حداقل رساندن آن رسمیت می یابد:

$$\hat{F} = \arg\min_{F} \mathbb{E}_{x,y}[L(y, F(x))]$$

روش تقویت گرادیان[59] یک y با مقدار حقیقی فرض می‌کند و به دنبال تقریبی $\hat{F}(x)$ در قالب مجموع وزنی توابع $h_i(x)$ از برخی از کلاس‌های $\mathcal{H}$، که یادگیرندگان پایه (یا ضعیف) نامیده می‌شوند:

$$\hat{F}(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + \text{const.}$$

معمولاً یک مجموعه آموزشی به ما داده می شود $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ از مقادیر نمونه شناخته شده x و مقادیر مربوط به y. مطابق با اصل تجربی کمینه‌سازی ریسک، این روش سعی می‌کند تقریبی $\hat{F}(x)$ را پیدا کند که میانگین مقدار تابع ضرر را در تمرین به حداقل برساند. مجموعه، یعنی ریسک تجربی را به حداقل می رساند. این کار را با شروع با یک مدل، متشکل از یک تابع ثابت $F_0(x)$ انجام می‌دهد و آن را به صورت حریصانه گسترش می‌دهد:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$$

$$F_m(x) = F_{m-1}(x) + \arg\min_{h_m \in \mathcal{H}} \left[\sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h_m(x_i)) \right]$$

که در آن $h_m \in \mathcal{H}$ یک تابع یادگیرنده پایه است.

متأسفانه، انتخاب بهترین تابع h در هر مرحله برای یک تابع از دست دادن دلخواه L به طور کلی یک مسئله بهینه‌سازی محاسباتی غیرممکن است. بنابراین، ما رویکرد خود را به یک نسخه ساده شده از مشکل محدود می کنیم.

ایده این است که شیب‌دارترین مرحله فرود را برای این مشکل کمینه‌سازی (نزول شیب عملکردی) اعمال کنیم.

ایده اصلی پشت پرشیب ترین فرود این است که با تکرار بر روی $F_m(x)$ حداقل محلی از تابع ضرر را پیدا کنید. در واقع، جهت حداکثر نزول محلی تابع تلفات، گرادیان منفی است.[10]

بنابراین، مقدار کمی γ را جابه‌جا می‌کنیم تا تقریب خطی معتبر باقی بماند:

$$F_m(x) = F_{m-1}(x) - \gamma \sum_{i=1}^{n} \nabla_{F_{m-1}} L(y_i, F_{m-1}(x_i))$$

جایی که $\gamma > 0$. این به معنی (برای γ کوچک: $L(y_i, F_m(x_i)) \leq L(y_i, F_{m-1}(x_i))$

Algorithm 1: Gradient Boosting

Data: training set $\{(x_i, y_i)\}_{i=1}^{n}$, a differentiable loss function $L(y, F(x))$, number of iterations M.
Result: $F_M(x)$.
Initialize model with a constant value:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma).$$

for $m \leftarrow 1$ **to** M **do**

- Compute pseudo-residuals:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n.$$

- Fit a base learner (or weak learner, e.g. tree) closed under scaling $h_m(x)$ to pseudo-residuals, i.e. train it using the training set $\{(x_i, r_{im})\}_{i=1}^{n}$
- Compute multiplier γ_m by solving the following one-dimensional optimization problem:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right).$$

- Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

end

6.1.3 روش پیشنهادی

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. [36] 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 [21]، پردازش زبان در ترانسفورماتورها، و پردازش داده‌های چندحسی (صدا، تصاویر، ویدئو، متن) در درک‌کننده‌ها است. [29, 44, 45, 50]

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 برچسپ و برچسب زده استفاده شده در طول آموزش‌ها بود. برای ساخت token ابتدا از روش مرسوم kmer در دی‌ان‌ای استفاده کردیم 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.

با توجه به پژوهش‌های انجام شده، به صورت جداگانه موفق به ارائه روشی مناسب برای حل مسئله نشده‌ایم، این امر به دلیل وجود نویز در داده به خاطر کم بودن ویژگی‌های مدل و تعداد کم داده‌های برچسب زده شده بود ولی با استفاده از کار پیشین و استفاده از تجربه آموزش مدل‌های دیگر روی تعداد داده بیشتر و ویژگی‌های بیشتر توانستیم روشی ارائه کنیم که عمومی‌تر و دقیق‌تر باشد.

فصل 4

نتایج شبیه‌سازی

در اینجا ما از داده‌های ارائه شده در مقاله DeepCRISPR برای مقایسه مدل‌های مختلف استفاده کرده‌ایم که حدود ۴۲۰ دنباله دی‌ان‌ای و نتیجه پیش‌بینی ۵ الگوریتم مختلف بود است، برای انجام آزمایش، ٪۸۰ داده‌ها را برای آموزش و ٪۲۰ داده‌ها را برای تست استفاده کرده‌ایم.

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	GAGTCGGGGGTTTCGTCATGTTGG	AGTAGAGTCGGGGGTTTCGTCATGTTGGTCA	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 1.4: هم بستگی اسپیرمن بین امتیاز موثر بودن و داده‌ها

1.0.4 Ensemble

نمونه‌ای از نتایج اولیه استفاده مستقیم روش‌های LPA و رگرسیون برای پیدا کردن وزن خوب بین اکسپرت‌ها با استفاده از کل داده‌ها با threshold های مختلف (کلاس‌بندی).

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 2.4: 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 3.4: 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 4.4: score F1

حال نتیجه روش پیشنهادی برای ادغام متدهای پیشین که با تقسیم ۸۰ به ۲۰ بدست آمده است و برای مقایسه رگرسیون آنها از رابطه اسپیرمن بین پیش‌بینی‌ها و داده واقعی و مربع تفاضلات میانگین استفاده کرده‌ایم. این روش را روی 100 تقسیم تصادفی امتحان کرده‌ایم و میانگین هر کدام را ارائه می‌دهیم:

Regerssion		
	Ours	DeepCRISPR
spearman_score	0.47784047	0.43845958
MSE_score	0.044421439	0.088784876
Classification		
Thershold = 0.7	Ours	DeepCRISPR
accuracy_score	0.63296875	0.53375
roc_auc_score	0.653567671	0.613415929
precision_score	0.792053405	0.847528917
recall_score	0.585135522	0.337152918
f1_score	0.671051272	0.480623603
Classification		
Thershold = 0.8	Ours	DeepCRISPR
accuracy_score	0.623203125	0.6003125
roc_auc_score	0.603931092	0.574306027
precision_score	0.67931511	0.694809806
recall_score	0.351809593	0.238997067
f1_score	0.46028489	0.353673895
Classification		
Thershold = 0.9	Ours	DeepCRISPR
accuracy_score	0.802578125	0.77734375
roc_auc_score	0.57223396	0.496869608
precision_score	0.450120453	0.16783153
recall_score	0.201583617	0.046421612
f1_score	0.271271654	0.070668875

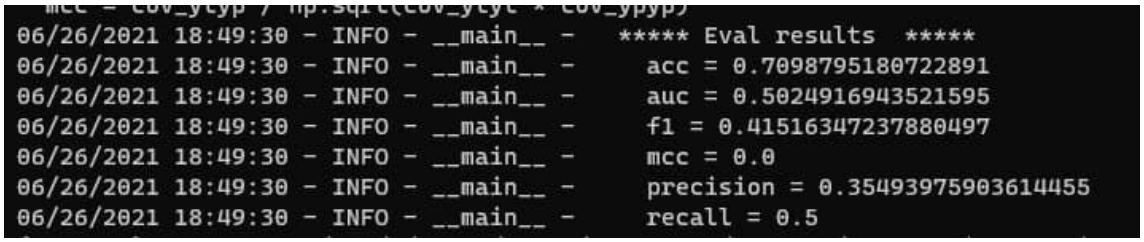
شکل 5.4: نتیجه آموزش

2.0.4 **Attention**

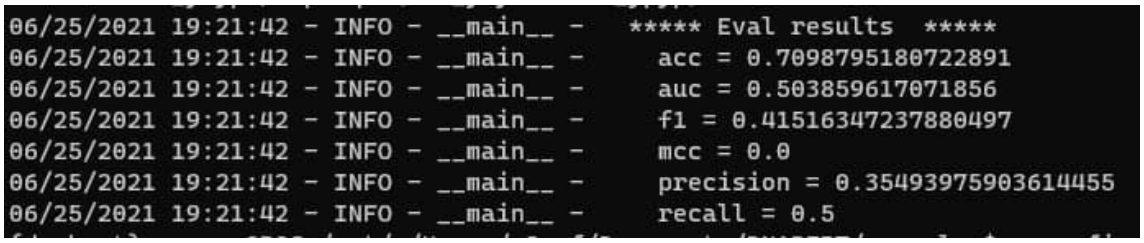
برای روش‌هایی که فقط از دنباله sgRNA استفاده می‌کنند، ابتدا حدود ۴ میلیون sgRNA از دادگان کارهای پیشین و ژن‌های مختلف جمع‌آوری کردیم و با کلمه‌ای و چندکلمه‌ای آن‌ها را توکنایزد کردیم، همچین از مدل از پیش‌آموزش شده روی DNA و مدل بدون آموزش قبلی برای آموزش مدل‌های bert استفاده کردیم و بردار بدست‌آمده را بروی دیتا با تقسیم ۸۰ به ۲۰ و threshold ۷.۰ کلاس‌بندی کردیم. 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 6.4: نتیجه تمرین به کمک ،3mer به کمک مدل DNAbert



شکل 7.4: نتیجه تمرین به کمک ،4mer به کمک مدل DNAbert



شکل 8.4: نتیجه تمرین به کمک ،6mer به کمک مدل DNAbert

نتیجه آموزش مدل توجه برای بدست‌آوردن بردار کد

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 9.4: نتیجه تمرین به کمک ،6mer به کمک مدل RoBerta

فصل 5

جمع‌بندی و کارهای آتی

دو مشکل اساسی که در داده‌ها پیدا می‌شود نویز ذاتی داده‌ها به خاطر حضور یک sgRNA در cell-line ها و ارگانیزم‌ها مختلف و نامتعادل بودن داده‌ها است چون معمولا کارشناسانی که sgRNA های مختلف را تست می‌کنند معمولا یک حس و بایاسی 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 بایاس‌ها در مدل باعث می‌شود که روشی جامع برای پیش‌بینی این تاثیرگذاری هاsgRNA بدست آید. نویز ذاتی داده‌ها نیز از این برگرفته می‌شود که یک sgRNA در یک ارگانیزم خاص می‌تواند خیلی خوب عمل کند ولی در ارگانیزم دیگر عمل‌کرد متوسط و یا ضعیفی داشته باشد و این که علت عمومی و جامعی برای چرایی موضوع پیدا نشده است و تمام ویژگی‌های بدست آمده حدودا حدس‌هایی است که با آزمایش‌ها پیدا شده است، در نتیجه امکان عمومی نبودن آنها بسیار بالاست. از جمله کارهایی که می‌توان برای حل این مشکل انجام داد این است که روش پیشنهادی به جای اینکه با ورودی متدهای دیگر پیاده‌سازی کنیم، روی ویژگی‌های بدست آمده پیاده سازی کنیم و مستقیما سعی به بهبود رگرسیون کنیم، البته این کار نیاز به مجموعه دادگان برزگی است که تمام ویژگی‌های مختلف پیدا شده را پوشش دهد، علاوه بر آن به نظر میرسد که تعداد ویژگی‌های پیدا شده بسیار بالاست در نتیجه باید بدنبال روشی برای انتخاب ویژگی‌های بهینه هم باشیم.

مراجع

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

[2] Labun, K., Montague, T. G., Krause, M., Torres Cleuren, Y. N., Tjeldnes, H., & Valen, E. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. doi:10.1093/nar/gkz365. (2019).

[3] T. G. Montague, J. M. Cruz, J. A. Gagnon, G. M. Church, E. Valen. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. doi:10.1093/nar/gku410. (2014).

[4] R. Jaenisch and B. Mintz. Simian Virus 40 DNA Sequences in DNA of Healthy Adult Mice Derived from Preimplantation Blastocysts Injected with Viral DNA. doi:10.1073/pnas.71.4.1250 (1974).

[5] A. M. Chakrabarty. Microorganisms having multiple compatible degradative energy-generating plasmids and preparation thereof. (1972).

[6] Kurzgesagt – In a Nutshell. Genetic Engineering Will Change Everything Forever – CRISPR. (2016). Retrieved 2021-06-06

[7] Wikipedia. https://en.wikipedia.org/wiki/CRISPR. Retrieved 2021-06-06

[8] addgene: All you need about CRISPR. https://www.addgene.org/guides/crispr/. Retrieved 2021-06-06

[9] A. Maxmen. Wired Easy DNA Editing Will Remake the World. Buckle Up. (2015) Retrieved 2021-06-06

[10] DW Zaharevitz, LW Anderson, Malinowski, Hyman, Strong, Cysyk. Contribution of de-novo and salvage synthesis to the uracil nucleotide pool in mouse tissues and tumors in vivo. (1992). doi:10.1111/j.1432-1033.1992.tb17420.x

[11] DNA: https://en.wikipedia.org/wiki/DNA. Retrieved 2022-01-14

[12] J. Craig Venter Institute. Genetics and Genomics Timeline (2004)

[13] glowing fish: https://www.glofish.com/. Retrieved 2022-01-14

[14] Patowary, K. Atomic Gardening: Breeding Plants With Gamma Radiation. (2013).

[15] Selective Breeding. https://en.wikipedia.org/wiki/Plant_breeding. Retrieved 2022-01-14

[16] Understanding DNA. https://medlineplus.gov/genetics/understanding/basics/dna/. Retrieved 2022-01-14

[17] Park, A. HIV Genes Have Been Cut Out of Live Animals Using CRISPR. (2016)

[18] Wendy Dong, B. Kantor. Lentiviral Vectors for Delivery of Gene-Editing Systems Based on CRISPR/Cas: Current State and Perspectives. doi:10.3390/v13071288 (2021).

[19] Building Blocks of the Genetic Code. https://www.ashg.org/discover-genetics/building-blocks/ (ed Figure 1: wikicommons) (2019). Retrieved 2022-01-16

[20] What is the Difference Between ZFN TALEN and CRISPR. https://www.differencebetween.com/what-is-the-difference-between-zfn-talen-and-crispr/ (ed Figure 01: ZFN) (2021). Retrieved 2022-01-16

[21] Alex Graves, G. W., Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwi□ska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, Adrià Puigdomènech Badia, Karl Moritz Hermann, Yori Zwols, Georg Ostrovski, Adam Cain, Helen King, Christopher Summerfield, Phil Blunsom, Koray Kavukcuoglu & Demis Hassabis Hybrid computing using a neural network with dynamic external memory. Nature 538 (7626), 471–476, doi:10.1038/nature20101 (2016).

[22] Allison, H. The Differences Between DNA and RNA (ed dna-versus-rna-sketch-Final.png) (2020). Retrieved 2022-01-16

[23] J. Doudna TED Talk: we can now edit our dna but let's do it Wisely https://www.ted.com/talks/jennifer_doudna_we_can_now_edit_our_dna_but_let_s_do_it_wisely/transcript?language=fa. (2015) Retrieved 2022-01-12

[24] Florian Heigwer, G. Kerr and M. Boutros E-CRISP: fast CRISPR target site identification. (2014).

[25] Bruening G., Lyons J. M. The case of the FLAVR SAVR tomato, https://calag.ucanr.edu/Archive/?article=ca.v054n04p6 (2000). Retrieved 2022-01-16

[26] Guohui Chuai, Qi Liu et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. (2018).

[27] Blockeel, H. Hypothesis Space. Encyclopedia of Machine Learning, 511–513, doi:10.1007/978-0-387-30164-8 373 (2011).

[28] 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 169, 5429–5433 (1987).

[29] Jaegle, A. G., Felix; Brock, Andrew; Zisserman, Andrew; Vinyals, Oriol; Carreira, Joao. Perceiver: General Perception with Iterative Attention. (2021).

[30] Jean-Paul Concordet, M. H. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Research 46, W242–W245, doi:10.1093/nar/gky354. (2018).

[31] Jeongbin Park, S. B., Jin-Soo Kim. Cas-Designer: a web-based tool for choice of CRISPR-Cas9 target sites. bioinformatics, doi:10.1093/bioinformatics/btv537. (2015).

[32] Johnson, I. S. Human insulin from recombinant DNA technology. science, doi:10.1126/science.6337396 (1983).

[33] Knoepfler, P. GMO Sapiens: The Life-Changing Science of Designer Babies. (2015).

[34] Kornel Labun, T. G. M., James A. Gagnon, Summer B. Thyme, Eivind Valen. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. (2016)

[35] 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, doi:10.1093/nar/gkx1268 (2017).

[36] Lecun, Y. Video lecture Week 6 of Deep Learning course at NYU (2020) Retrieved 2021-12-13.

[37] Ledford, H. CRISPR: gene editing is just the beginning. nature 531, 156–159 (2016).

[38] Ledford, H. HIV cut from cells and rats with CRISPR. nature 531, pages156–159 (2016).

[39] Mojica, F. J., Juez, G. & Rodriguez-Valera, F. Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified PstI sites. Molecular Microbiology 9, 613–621 (1993).

[40] Opitz, D. M., R. Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research 11, 169–198, doi:10.1613/jair.614 (1999).

[41] Patrick D. Hsu, E. S. L., and Feng Zhang. Development and Applications of CRISPR-Cas9 for Genome Engineering, 2014).

[42] Polikar, R. Ensemble based systems in decision making. IEEE Circuits and Systems Magazine 6 (3), 21–45, doi:10.1109/MCAS.2006.1688199 (2006).

[43] Rafal Kaminski, Y. C., Tracy Fischer, Ellen Tedaldi, Alessandro Napoli, Yonggang Zhang, Jonathan Karn, Wenhui Hu & Kamel Khalili. Elimination of HIV-1 Genomes from Human T-lymphoid Cells by CRISPR/Cas9 Gene Editing. Scientific Reports 6 (2016).

[44] Ramachandran, P. P., Niki; Vaswani, Ashish; Bello, Irwan; Levskaya, Anselm; Shlens, Jonathon. Stand-Alone Self-Attention in Vision Models. (2019).

[45] Ray, T. Google's Supermodel: DeepMind Perceiver is a step on the road to an AI machine that could process anything and everything. ZDNet (2021).

[46] Reardon, S. First CRISPR clinical trial gets green light from US panel. Nature (2016).

[47] Rokach, L. Ensemble-based classifiers. Artificial Intelligence Review 33, 1–39, doi:10.1007/s10462-009-9124-7 (2010).

[48] Sangsu Bae 1, J. P., Jin-Soo Kim. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. bioinformatics, doi:10.1093/bioinformatics/btu048.

[49] 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, doi:10.1371/journal.pone.0124633 (2015).

[50] Vaswani, A. S., Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz; Polosukhin, Illia Attention Is All You Need. (2017).

[51] Walsh, G. Therapeutic insulins and their large-scale manufacture. doi:10.1007/s00253-004-1809-x (2005).

[52] Wong, J. L. a. K.-C. Off-target predictions in CRISPR-Cas9 gene editing using deep learning. Bioinformatics, doi:10.1093/bioinformatics/bty554 (2018).

[53] Thomas Gaj, Charles A. Gersbach, and Carlos F. Barbas. ZFN, TALEN and CRISPR/Cas-based methods for genome engineering. doi:10.1016/j.tibtech.2013.04.004 (2014).

[54] Alvaro L. Pérez-Quintero, L. M. Rodriguez-R, A. Dereeper, C. López, R. Koebnik, B. Szurek, and S. Cunnac. An Improved Method for TAL Effectors DNA-Binding Sites Prediction Reveals Functional Convergence in TAL Repertoires of Xanthomonas oryzae Strains. doi: 10.1371/journal.pone.0068464 (2013).

[55] David H.Wolpert. Stacked generalization. doi:10.1016/S0893-6080(05)80023-1 (1992)

[56] Chaya Bakshi. Random Forest Regression Picture. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84 (2020). Retrieved 2022-03-23

[57] Leo Breiman. Random Forest. doi:10.1023/A:1010933404324 (2001).

[58] Pierre Geurts, D. Ernst, L. Wehenkel. Extremely randomized trees. doi:10.1007/s10994-006-6226-1 (2006).

[59] Jerome H. Friedman. Stochastic Gradient Boosting. doi:10.1016/S0167-9473(01)00065-2 (2002).

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

March 31, 2022