



دانشگاه صنعتی شریف
دانشکده علوم ریاضی

پایان نامه کارشناسی ارشد
ریاضی کاربردی

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

نگارش
محمد رستمی

استاد راهنما
دکتر محسن شریفی تبار

استاد راهنما دوم
دکتر حمیدرضا ربیعی

استاد مشاور
دکتر محمدحسین رهبان

17 فروردین 1401

به نام او
دانشگاه صنعتی شریف
دانشکده علوم ریاضی

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

کمیته داوران

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

قدردانی

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

و تشکر از خداوند.

چکیده

تناوب‌های کوتاه پالیندروم فاصله‌دار منظم خوش‌های یا به طور خلاصه، کریسپر (CRISPR) یکی از روش‌های نوین است که متخصصان ژنتیک و محققان پژوهشگرانی را قادر می‌سازد تا با حذف بخش‌هایی از ژنوم، افزودن یا تغییر بخش‌هایی از آن در دنیان ای (DNA) تغییر ایجاد کنند. این فناوری، نوعی سیستم ایمنی تطابق‌پذیر در باکتری‌ها است که با کمک آن می‌توان بسیاری از بیماری‌ها مانند نایینواری و ناشنواری و حتی سرطان را درمان کرد. یکی از مشکلات بزرگ در استفاده موفق از کریسپر، پیش‌بینی دقیق تاثیر راهنمای آران‌ای (Guide RNA) روی هدف و حساسیت خارج از هدف است. در حالی که برخی از روش‌ها این طرح‌ها را طبقه‌بندی می‌کنند، بیشتر الگوریتم‌ها بر روی داده‌های جداگانه بازن‌ها و سلول‌های مختلف هستند. عدم تعمیم این روش‌ها مانع استفاده از این راهنمای در آزمایشات بالینی می‌شود، زیرا برای هر درمان، این فرایند باید دقیقاً برای همان سلول درست شده باشد و عموماً داده‌کافی برای طراحی الگوریتم در آن سلول در دسترس نیست. در این پژوهش، روشی پایدار برای ادغام نتایج روش‌های مختلف برای تخمین دقیق تأثیرگذاری یک راهنمای ارائه می‌دهیم. از آنجایی که این روش با تعداد داده‌ی کمی، دارای دقت بالایی است، روشی مناسب برای استفاده در مسئله‌هایی است که تعداد داده بسیار کم است.

فهرست مطالب

| | | | |
|----|--|-------|---|
| 1 | | مقدمه | 1 |
| 1 | نوکلئوتید | 1.1 | |
| 1 | آران‌ای | 2.1 | |
| 2 | دیان‌ای | 3.1 | |
| 2 | تفاوت‌های دیان‌ای و آران‌ای | 1.3.1 | |
| 3 | ویرایش ژنوم | 4.1 | |
| 3 | شکست و تعمیر دیان‌ای | 1.4.1 | |
| 4 | Zinc finger nucleases (ZFN) | 2.4.1 | |
| 4 | TALEN | 3.4.1 | |
| 5 | کریسپر | 5.1 | |
| 5 | کریسپر در باکتری | 1.5.1 | |
| 6 | عمل کرد کریسپر در ژن | 2.5.1 | |
| 6 | حساسیت | 3.5.1 | |
| 7 | تأثیرگذاری | 4.5.1 | |
| 7 | انواع کریسپر | 5.5.1 | |
| 10 | کارهای پیشین | 2 | |
| 10 | روش‌های مستقیم | 1.2 | |
| 10 | [34,3,2] Chopchop | 1.1.2 | |
| 10 | [48,31] Cas-Designer و Cas-OFFinder | 2.1.2 | |
| 11 | [24] E-CRISP | 3.1.2 | |
| 12 | [30] CRISPOR | 4.1.2 | |
| 12 | روش‌های یادگیری ژرف | 2.2 | |
| 12 | پیش‌بینی off-target به کمک یادگیری ژرف | 1.2.2 | |
| 13 | [49] CCTop | 2.2.2 | |
| 13 | DeepCRISPR | 3.2 | |
| 15 | روش‌های پیشنهادی | 3 | |
| 16 | Learning Ensemble | 1.3 | |
| 16 | تعریف | 1.1.3 | |
| 16 | رگرسیون با جنگل تصادفی | 2.1.3 | |
| 17 | درختان بسیار تصادفی | 3.1.3 | |
| 18 | حدائق مربعات معمولی | 4.1.3 | |
| 18 | تقویت گرادیان | 5.1.3 | |
| 19 | روش پیشنهادی | 6.1.3 | |
| 21 | آنالیز مشخصات پنهان | 7.1.3 | |
| 22 | Attention | 2.3 | |
| 24 | نتایج شبیه‌سازی | 4 | |
| 25 | LPA | 1.4 | |

| | | |
|----|-----------------------|-------|
| 26 | روش پیشنهادی | 2.4 |
| 27 | Attention | 1.2.4 |
| 28 | جمع‌بندی و کارهای آتی | 5 |

فهرست تصاویر

| | | |
|----|---|-----|
| 1 | یک حلقه از pre-mRNA نوکلئوتیدها (سبز) و ستون فقرات ریبوز فسفات (آبی) مشخص شده اند. این یک رشته منفرد از آران ای است که بر روی خود تا می شود. عکس گرفته شده از ویکی پدیا | 1.1 |
| 2 | شکل دو بعدی دی ان ای [19] | 2.1 |
| 2 | مقایسه دی ان ای و آران ای [22] | 3.1 |
| 3 | مکانیزم ترمیم دی ان ای، عکس گرفته شده از ویکی پدیا | 4.1 |
| 4 | مکانیزم TALEN [20] | 5.1 |
| 5 | مکانیزم ساده ای از CRISPR [18] | 6.1 |
| 6 | مکانیزم CRISPR [8] | 7.1 |
| 7 | مکانیزم TALEN [8] | 8.1 |
| 11 | (الف) شماتیک مکان های دی ان ای یا آران ای نشان می دهد. (ب) استراتژی برای برآمدگی 1-nt DNA یا آران ای 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 دی ان ای (وسط) یا آران ای (پایین) ارائه می شود. علاوه بر این، کاربر می تواند موارد مربوطه را به دست آورد اطلاعات زنومی از طریق مرورگر ژنوم Ensembl و همکاران، 2011)، با کلیک بر روی دکمه "اطلاعات در Flicek [48]" | 1.2 |
| 11 | الگوریتم E-CRISP [24] | 2.2 |
| 12 | هم بستگی اسپیرمن بین امتیاز موثر بودن و داده ها [30] | 3.2 |
| 12 | هم بستگی اسپیرمن بین امتیاز موثر بودن و داده ها [52] | 4.2 |
| 13 | [26] DeepCRISPR | 5.2 |
| 21 | شمای روش پیشنهادی | 1.3 |
| 21 | شمای دقیق استفاده شده برای بدست آوردن نتیجه | 2.3 |
| 23 | مدل دقت DNAAbert [60] | 3.3 |
| 25 | پراکندگی داده نسبت به کلاس ها | 1.4 |

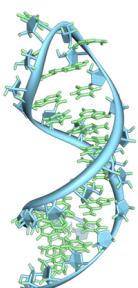
فصل 1

مقدمه

مقیاس در حال گسترش و پیچیدگی ذاتی داده‌های بیولوژیکی، استفاده روزافزون از یادگیری ماشین در زیست‌شناسی را برای ساختن مدل‌های آموزنده و پیش‌بینی کننده فرآیندهای بیولوژیکی اساسی تشویق کرده است. در ویرایش ژن‌ها نیز، این روش‌ها موثر هستند زیرا تعداد عوامل موثر در موقیت ویرایش ژن (تائیرگذاری) بسیار بالا و نقش هر کدام از ویژگی‌ها مهم است، علاوه بر آن بدست آوردن تمام این عوامل، پیچیده و هزینه بر است و همچین، پیش‌بینی اثرات بوجود آمده و مناطق تغییر کرده و ناخواسته (حساسیت) کاری سخت و تصادفی است که برای مدل‌های یادگیری ماشین امری مرسوم است. عموماً روش‌های ویرایش ژن‌ها، امری هزینه بر با داده‌های کم است ولی با پیشرفت علم، روشی مناسب و کم هزینه به نام کریپسپر برای ویرایش ژن بدست آمده است ولی قبل از این که با کریپسپر آشنا شویم، بهتر است راجع به تاریخچه ویرایش ژن‌ها صحبت کنیم. انسان‌ها سال‌هاست که مشغول به ویرایش و مهندسی ژن‌ها هستند، با استفاده از پروژه انتخابی^۱. اصلاحات نژادی متعددی در گیاهان و حیوانات مخصوصاً گونه‌های کلیدی مانند گندم، برنج و سگ‌ها ایجاد شده است. انسان‌ها در این کار شدیداً ماهر شده‌اند به طوری که در صده گذشته، تعداد دانه‌های هر شاخه گندم چندین برابر و ارتفاع آنها کوتاه‌تر شده است تا در معرض خطر کمتری باشند و حدود ۸۰ نژاد جدید سگ به وجود آمده است. البته با وجود پیشرفت‌های متعدد انسان‌ها، تا قبل از کشف دی‌ان‌ای، انسان‌ها ساز و کار دقیق آن را نمی‌دانستند.

1.1 نوکلئوتید

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



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

نوکلئوتیدها از سه زیر واحد مولکولی تشکیل شده‌اند: یک باز نوکلئوتیدی، یک قند پنج کربن‌هه پنتوز (ریبوز یا دئوكسی ریبوز)، و یک گروه فسفات شامل یک تا سه فسفات. چهار باز نوکلئوتیدی دی‌ان‌ای شامل: گوانین، آدنین، سیتوزین و تیمین می‌باشند؛ در آران‌ای، اوراسیل به جای تیمین استفاده می‌گردد.

2.1 آران‌ای

اسید ریبونوکلئیک^۲ 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 G, C و A U، مشخص می‌شوند) است که برخودش پیچ می‌خورد، برخلاف 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

¹Selective Breeding

²RiboNucleic Acid

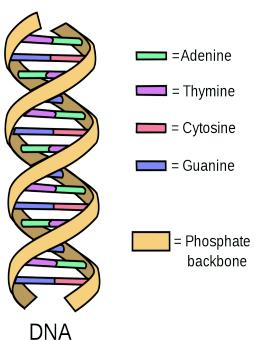
می‌کند، آران‌ای پیک یا پیامبر (mRNA) می‌گویند. نوعی دیگر از آران‌ای، آران‌ای حامل (tRNA) است که اسیدهای آمینه را به ریبوزوم منتقل می‌کند، تا ریبوزوم، اسیدهای آمینه را بر اساس اطلاعات موجود در mRNA کنار یک دیگر قرار دهد. نوع دیگر، آران‌ای ریبوزومی (rRNA) است که در ساختار ریبوزوم‌ها شرکت دارد؛ این موضوع به این معناست که ریبوزوم (رناتن) ها متشكل از پروتئین‌ها و آران‌ای‌های ریبوزومی هستند.

3.1 دیان‌ای

دئوکسی ریبو نوکلئیک اسید³ به اختصار دیان‌ای یک مولکول متشكل از دو زنجیره پلی نوکلئوتیدی است که به دور یکدیگر می‌پیچند و دارای دستورالعمل‌های ژنتیکی است که برای کارکرد و توسعه زیستی جانداران و ویروس‌ها مورد استفاده قرار می‌گیرد. نقش اصلی مولکول دیان‌ای ذخیره‌سازی طولانی مدت اطلاعات ژنتیکی و دستوری است. لیپیدها، پروتئین‌ها، کربوهیدرات‌های 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 می‌شوند زیرا از واحدهای مونومر یا تکپار ساده‌تری به نام نوکلئوتید تشکیل شده‌اند. هر نوکلئوتید از یکی از چهار نوکلئوباز حاوی نیتروژن (سیتوزین، C، گوانین، G، آدنین، A یا تیمین T)، کربوهیدرات‌پنج‌کربنیه به نام دئوکسی ریبوز و یک گروه فسفات تشکیل شده است. نوکلئوتیدهای دیگر زنجیره، توسط پیوندهای کووالانسی (معروف به پیوند فسفودی‌استر) بین قند یک نوکلئوتید و فسفات نوکلئوتید بعدی به یکدیگر متصل می‌شوند و در نتیجه یک ستون فقرات قند-فسفات متناوب ایجاد می‌شود.

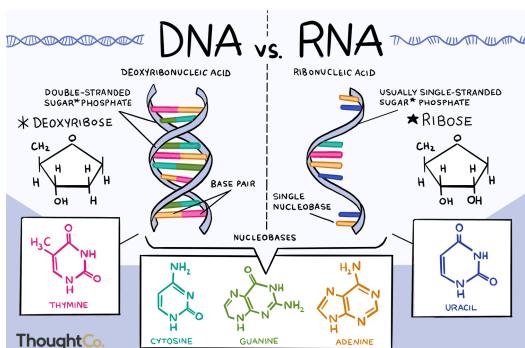
با زهای نیتروژنی، دو رشته پلی نوکلئوتیدی جداگانه، طبق قوانین جفت شدن بازها (A با T و C با G)، با پیوندهای هیدروژنی به یکدیگر متصل می‌شوند تا دیان‌ای دو رشته‌ای بسازند. این دو رشته مکمل، ناهمسو و محلول (در آب) هستند (دیان‌ای حلقوی قطبیت ندارد اما هر رشته از دیان‌ای خطی دارای قطبیت است). بازهای نیتروژنی مکمل به دو گروه پیریمیدین‌ها و پورین‌ها تقسیم می‌شوند. در دیان‌ای، پیریمیدین‌ها 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 می‌کند. این اطلاعات زمانی که دو رشته از هم جدا می‌شوند، تکرار می‌شوند. بخش بزرگی از دیان‌ای (بیش از 98٪ برای انسان) کد نشده⁴ 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 آینین نامگذاری ترکیب‌های شیمیایی، انتهایی کردن در حلقة شکری نوکلئوتید شماره گذاری شده‌اند. هر رشته دیان‌ای یا آران‌ای دارای یک پایانه⁵ که معمولاً شامل یک گروه فسفاتی است و یک پایانه³ که معمولاً از جانشین ریبوز اصلاح شده OH- است. به هر قند یکی از چهار نوع نوکلئوباز (یا باز) متصل است.

توالی این چهار هسته در امتداد ستون فقرات است که اطلاعات ژنتیکی را رمزگذاری می‌کند. رشته‌های آران‌ای با استفاده از رشته‌های دیان‌ای به عنوان یک الگو در فرآیندی به نام رونویسی ایجاد می‌شوند که در آن بازهای دیان‌ای 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.



شکل 1.3: مقایسه دیان‌ای و آران‌ای [22]

1.3.1 تفاوت‌های دیان‌ای و آران‌ای

تفاوت‌ها:

- دیان‌ای بر عکس آران‌ای از هسته سلول خارج نمی‌شود.
- آران‌ای بدون زن می‌باشد.
- دیان‌ای در ذخیره و آران‌ای در انتقال اطلاعات و راثتی و در ساختار ریبوزوم نقش دارد.
- مولکول دیان‌ای دور رشته‌ای در هم تنیده اما مولکول آران‌ای تک رشته‌ای است.

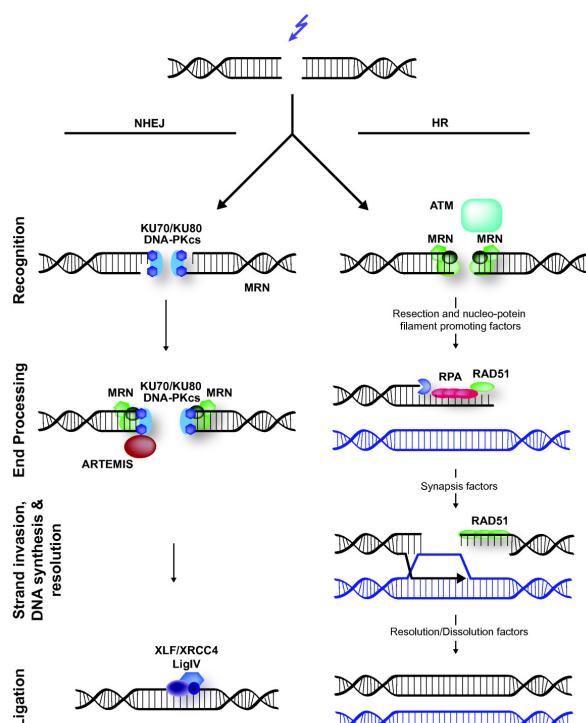
³Deoxyribonucleic acid

⁴non-coding

- در دی‌ان‌ای باز آلی یوراسیل و در آران‌ای باز آلی تیمین شرکت ندارد (U در دی‌ان‌ای و T در آران‌ای).
- قند پنج کربنه موجود در دی‌ان‌ای را دئوکسی ریبوز و در آران‌ای قند ریبوز نامیده می‌شود. تفاوت بین قندها وجود گروه هیدروکسیل بر روی کربن ۲ ریبوز و عدم وجود آن در کربن ۲ دئوکسی ریبوز است.

شهاهت‌ها:

- هر دو پلیمر هستند و از نوکلئوتید تشکیل شده‌اند.
- در هر دو نوکلئوتیدهای مقابله 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 ۴.۱: مکانیزم ترمیم دی‌ان‌ای، عکس گرفته شده از ویکی‌پدیا.

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

1.4.1 شکست و تعمیر دی‌ان‌ای

NHEJ

پرتوهای یونیزه کننده و برخی داروهای ضد سرطان باعث شکست هر دو رشته‌ی دی‌ان‌ای می‌شوند. سیستمی که برای ترمیم این نوع آسیب به کار گرفته می‌شود، سیستم ترمیم اتصال انتهای غیر همولوگ (NHEJ) می‌باشد که مستعد به خطابه شمار می‌رود، زیرا همواره چندین نوکلئوتید در جایگاه ترمیم از بین می‌روند و دو انتهای شکسته شده از کروموزوم‌های همولوگ یا غیر همولوگ به یکدیگر متصل

⁵Zinc Finger Nuclease

⁶Transcription activator-like effector nuclease

⁷Clustered Regularly Interspaced Short Palindromic Repeats

⁸Double-Strand Break (Cut)

⁹Non-Homologous End Joining

¹⁰Homology Directed Repair

می‌شوند. زمانی که کروماتیدهای خواهی برای ترمیم شکسته‌های 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 می‌یابند، آن‌گاه در هر انتها چندین باز توسط نوکثاز حذف شده و دو مولکول از طریق آنزیم لیگاز به هم متصل می‌گردند. DSB‌ها ترجیحاً در سلول، توسط اتصال انتها یک غیر همولوگ (NHEJ) ترمیم می‌شوند، مکانیزم سریعی که اغلب باعث درج یا حذف (indels) در دی‌ان‌ای می‌شود. ایندل‌ها اغلب منجر به تغییر اساسی در دی‌ان‌ای می‌شوند، به‌طوری که دی‌ان‌ای عملکرد خود را از دست می‌دهند. پس در نتیجه معمولاً به عنوان سلول مرده در نظر گرفته می‌شوند و حذف می‌شوند. برای ویرایش ژنوم مطمئن، اعمال شدن ویرایش و تغییر نکردن آن نکات مهمی است. پس دانشمندان تمام تلاش‌شان را می‌کنند که بعد از DBS، دی‌ان‌ای به این روش تعمیر نشود.

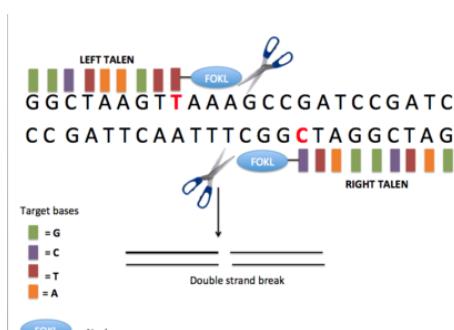
HDR

تعمیر هدایت شده همولوژی (HDR)، مکانیزمی در سلول‌ها برای ترمیم ضایعات دی‌ان‌ای دو رشته‌ای، از یک نسخه مشابه دی‌ان‌ای برای ترمیم استفاده می‌کند. رایج‌ترین شکل، HDR نوترکیبی همولوگ است. مکانیسم HDR تنها زمانی می‌تواند 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 برنامه‌ریزی با کاربرد وسیع است. زنجیره ای از پروتئین‌های انگشت روی هستند که به یک نوکلئاز باکتریایی ملحق شده اند تا بتوانند سیستمی را تولید کنند که قادر به ایجاد برش‌های دو رشته ای خاص در دی‌ان‌ای برای ویرایش ژن باشد. پروتئین‌های انگشت روی هدف قرار دادن ناحیه خاص را فراهم می‌کنند زیرا هر یک از آنها سه جفت باز یا 3bp از دی‌ان‌ای را شناسایی می‌کنند. نوکلئازی که معمولاً در تکنولوژی ZFN متصل به زنجیره پروتئین‌های 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 می‌گیرد. این آنزیم‌ها کمک زیادی به تولید موجودات ترانس‌ژنیک می‌کنند و بدليل اینکه فراوانی نوترکیبی همولوگ بسیار ناچیز بوده، اهمیت زیادی در مهندسی ژنتیک و مطالعات ترانس‌ژنیک، ناک اوت و غیره پیدا کرده‌اند. این پروتئین‌های مهندسی شده و متصل شونده به دی‌ان‌ای می‌توانند ژنوم را در جایگاه‌های ویژه‌ای شناسایی کرده و ایجاد برش‌های دو رشته‌ای کنند. در صورتی که سیستم تعمیر NHEJ فعال شود چون این سیستم ترمیم مستعد خطاست، سبب ایجاد جهش در آن ناحیه خاص از ژنوم می‌شود بنابراین در مطالعات موتاژنر نیز اهمیت دارند. انتقال یک بردار حاوی ژن مورد نظر به همراه ZFNs سبب تسهیل درج ژن در آن ناحیه از ژنوم می‌گردد.

TALEN 3.4.1



شکل ۵.۱: مکانیزم [20] TALEN

نوکلئازهای رونویس مؤثر-مانند فعال کننده (TALENs)¹¹ پروتئین‌های متصل شونده به دی‌ان‌ای با آرایه تکراری 33 یا 34 اسید آمینه هستند. TALEN‌ها آنزیم‌های محدود کننده مصنوعی هستند که از ادغام حوزه برش دی‌ان‌ای یک نوکلئاز با دامنه‌های طراحی شده اند، که می‌توانند به طور خاص یک توالی دی‌ان‌ای منحصر به فرد را شناسایی کنند. این پروتئین‌های ادغام شده به عنوان قیچی دی‌ان‌ای برای ویرایش یک ژن خاص عمل می‌کنند که به راحتی قابل برنامه‌نویسی بوده و 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.

TALE‌هایی که رشته ای از آمینواسیدها هستند که هر کدام وظیفه دارند که یک تک نوکلئوتید را پیدا کنند. نوکلئاز می‌تواند شکستگی‌های دو رشته‌ای را در محل هدف ایجاد کند، که قادر است

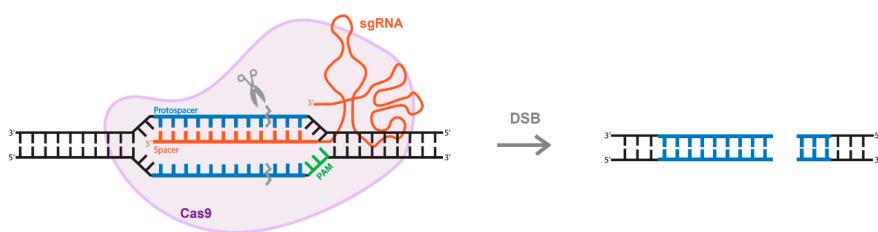
با اتصال انتها یک غیر همولوگ (NHEJ) ترمیم شود، که منجر به اختلالات ژنی از طریق وارد کردن یا حذف‌های تک نوکلوتیدی می‌شود، هر تکرار حفظ می‌شود. به استثنای دی‌باقیمانده متغیر تکرار شونده (RVDs)،¹² که در موقعیت‌های آمینواسید 12 و 13 قرار دارند. RVD‌ها توالی دی‌ان‌ای را تعیین می‌کنند که TALE به آن متصل می‌شود. 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 می‌توان با کاتالیزوری از یک نوکلئاز از دی‌ان‌ای به نام FokI، ادغام کرد تا با آن‌ها TALEN را ساخت. ساختارهای TALEN توالی‌های

Transcription activator-like effector nucleases¹¹
Repeat Variable Di-residues¹²

دی‌ان‌ای را فقط در مکان‌های از پیش انتخاب شده متصل می‌کنند و می‌شکنند. هدف TALEN را می‌توان 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 ۶ جفت نوکلئوتید باشد، TALEN می‌تواند آن را ویرایش کند. سازه‌های TALEN به روشهای مشابه با نوکلئازهای انگشت روی طراحی شده استفاده می‌شوند و دارای سه مزیت در جهش زایی هدفمند هستند: (۱) احتمال ویرایش درست هدف بالاتر است، (۲) اثرات خارج از هدف کمتر است و (۳) طراحی آن آسان‌تر است.^[54]

5.1 کریسپر

Clustered Regularly Interspaced Short Palindromic Repeats کوتاه پالیندرومِ فاصله‌دار منظم خوش‌های به اختصار کریسپر (به انگلیسی: CRISPR) یا به معنی "تناوب‌های بخشی از دی‌ان‌ای پروکاریوت هستند که حاوی تناوب‌های کوتاه توالی‌های بنیادین هستند. این 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 معمیوب 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 سلطان 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 "کس 9" (Cas9)^[23] هست. این آنزیم به عنوان یک جفت قیچی مولکولی می‌تواند دو رشته دی‌ان‌ای را در محل خاصی از ژنوم برش دهد.^[23]



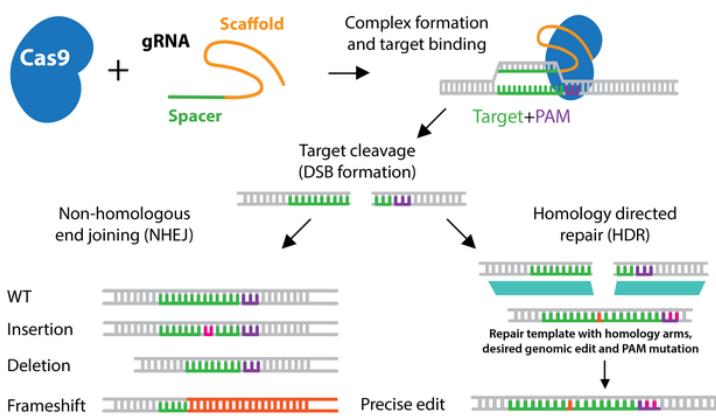
شکل 6.1: مکانیزم ساده شده‌ای از CRISPR^[18]

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. Francisco Mojica [28] و همکارانش در سال ۱۹۹۳ تکرارهای مشابهی را در چندین گونه میکروبی دیگر یافته‌اند.^[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 روش‌ها بین تکرارهای مستقیم تقسیم شده‌اند که اجازه می‌دهند سیستم CRISPR، به طور ایمن و دقیق و نه به طور غیر ایمن، شناسایی شود. آرایه 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: بیوژن و پردازش CRISPR) در نوع I و III، رونوشت pre-CrRNA توسط ریبونوکلئازهای مرتبط با CRISPR، شکسته می‌شوند و این کار موجب آزاد شدن چندین CrRNAs کوچک می‌شود. به طور متوسط CrRNA نوع III بیشتر در انتهای ۳' توسط RNase Hایی که هنوز مشخص نشده‌اند برای تولید رونوشت کاملاً بالغ پردازش می‌شوند. در نوع II، یک آران‌ای کریسپر فعال کننده ترانس است که با تکرارهای مستقیم هیبرید می‌شود و یک آران‌ای دوپلکس را تشکیل می‌دهد و توسط RNase III درونی و نوکلئازهای tracrRNA (tracrRNA) که با تکرارهای CRISPR های بالغ شده نوع I و III سیستم CrRNA را شکسته و پردازش می‌شود. CRISPR ناشناخته دیگر شکسته و پردازش می‌شود. CRISPR های بالغ شده نوع I و III کمپلکس‌های Cas9 پروتئینی برای تشخیص و تحریب توالی هدف اضافه می‌شوند. در سیستم‌های نوع II، کمپلکس هیبرید CRISPR-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 تنظیم کننده برای تسهیل شناسایی توالی هدف استفاده می‌کنند. در CRISPR نوع I، کمپلکس Cascade با یک مولکول Cascade به Cas9 متصل شده که یک مجموعه نظارتی بی نظیری است که دی‌ان‌ای هدف را شناسایی می‌کند. سپس نوکلئاز Cas3، لوب Cas3 به کار گرفته و به آن متصل می‌شود و واسطه تحریب توالی هدف می‌شود. در CRISPR نوع III، سیستم مداخله گر تنظیم کمپلکس‌های Cmr به ترتیب متصل شده و به ترتیب سوبیسترها را می‌شناسند. در مقابل، سیستم نوع II فقط نیاز به Csm یا به RNA را می‌شکنند. برای تحریب دی‌ان‌ای جفت شده با آران‌ای راهنمای دوپلکس خود دارد که این آران‌ای راهنمای حاوی ترکیبی از CRISPR-tracrRNA است.

2.5.1 عمل کرد کریسپر در ژن

همانطور که گفتیم، مدل‌های مختلفی از CRISPR را به حال درست شده است ولی به صورت کلی می‌توان CRISPR را به دو قسمت آران‌ای و cas تقسیم کرد که در آن وظیفه جدا کردن دو رشته دی‌ان‌ای را از هم دارد و آران‌ای که هدف را مشخص و قیچی می‌کند. برای این که دقیقاً نقطه شکست دی‌ان‌ای مشخص شود، cas نیاز به یک علامت دارد که با رسیدن به آن کار خود را شروع کند. به این رشته گفته می‌شود که کاملاً 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 گویند، استفاده کرده‌ایم. این طرح باعث محدود شدن هدف‌های مورد استفاده می‌شود به طوری که باشد که در آن N یک نوکلیوتید دلخواه است. در نتیجه رشته هدف همیشه با NGG ختم می‌شود.



[8] مکانیزم CRISPR

3.5.1 حساسیت

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

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

بنابراین، دو نیکاز که رشته‌های دی‌ان‌ای مخالف را هدف قرار می‌دهند، 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 می‌تواند با ویرایش ژن با واسطه HDR برای ویرایش‌های ژنی خاص ترکیب شود.

در سال 2015، محققان از rational mutagenesis برای توسعه دو Cas9 با ثبات بالا استفاده کردند: eSpCas9 و SpCas9-HF1. eSpCas9 برای توسعه دو Cas9 با آلتین این است که برهمنکنش‌های بین شیار HNH/RuvC و رشته دی‌ان‌ای غیرهدف را نضعیف 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 آلتین کاهش می‌دهد که برهمنکنش Cas9 با ستون فقرات فسفات دی‌ان‌ای را مختلط می‌کند. یکی دیگر از Cas9 با حساسیت 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 آنژیم با حساسیت بالا نسبت به Cas9 نوع وحشی، ویرایش خارج از هدف کمتری تولید می‌کنند.

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

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

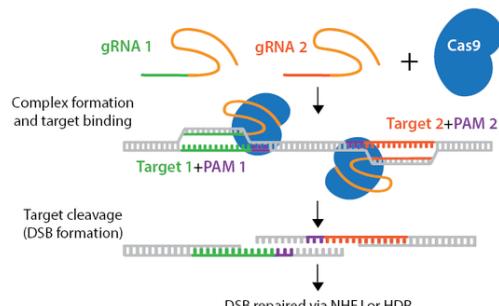
هر دو نوع ویرایشگر پایه با چندین نوع Cas9 از جمله Cas9 با ثبات بالا در دسترس هستند. پیشرفت‌های بیشتری با بهینه‌سازی بیان پروتئین، اصلاح ناحیه پیوندی بین نوع Cas و دامیناز برای تنظیم پنجره ویرایش، یا افزودن ترکیب‌هایی که خلوص محصول را افزایش می‌دهند مانند مهارکننده دی‌ان‌ای گلیکوزیلاز (UGI) یا Gam مشتق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

5.5.1 انواع کریسپر

طبقه‌بندی rovA و همکاران ۵ نوع سیستم کریسپر را تعریف می‌کند که دارای ۱۶ زیرنوع بر اساس ویژگی‌های مشترک و شباهت تکاملی است. که به دو دسته بزرگ تقسیم می‌شوند. کلاس‌ها بر اساس ساختار پیچیده‌ای است که دی‌ان‌ای ژنوم را تجزیه می‌کند. نوع II 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



شکل ۸.1: مکانیزم [8] TALEN

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) ویرایش (hypermutation) را در یک پنجره ویرایش گسترده‌تر ایجاد می‌کنند و بنابراین برای تکامل هدایت شده مناسب هستند. نمونه‌هایی از این سیستم‌های ویرایش پایه عبارتند از جهش‌زایی هدفمند با واسطه AID (TAM) و CRISPR-X (AID)، که در آن Cas9 با سیتیدین دامیناز (AID) ناشی از فعل سازی ترکیب می‌شود.

نیکاز CRISPR/Cas جهش‌بافته، به جای شکستگی‌های دو رشتہ‌ای ایجاد شده توسط آنزیم‌های Cas، شکستگی‌های تک رشتہ‌ای با هدف gRNA را در دی‌ان‌ای ایجاد می‌کنند. برای استفاده از جهش نیکاز، به دو gRNA نیاز دارید که رشتہ‌های مخالف دی‌ان‌ای شما را در مجاورت یکدیگر مورد هدف قرار دهند. این شیارهای دوتایی یک شکست دو رشتہ‌ای (DSB) ایجاد می‌کنند که با استفاده از اتصال انتهایی غیر همولوگ (NHEJ) و مستعد خطا تعمیر می‌شود. استراتژی‌های دوتایی اثرات ناخواسته off-targets را کاهش می‌دهند. جهش‌بافته‌های نیکاز همچنین می‌توانند با یک الگوی تعمیر برای معرفی ویرایش‌های خاص از طریق تعمیر هدایت شده همولوژی (HDR) استفاده شوند.

در حالی که SpCas9 (S. pyogenes Cas9) 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 هدایت شده همولوژی (HDR)، که نیاز به توالی‌های PAM در مجاورت بسیار نزدیک به منطقه برای ویرایش را دارد، نگران کننده است.

برای رسیدگی به این محدودیت‌ها، محققان آنژیم‌های SpCas9 را با ویژگی‌های تغییر یافته PAM با استفاده از روش‌های مختلفی از جمله تکامل به کمک فاژ و جهش‌زایی هدایت شده مهندسی کرده‌اند. این منجر به توسعه چندین نوع مشتق شده از SpCas9 با توالی

های PAM غیر NGG شد. جایگزین دیگر Cas9x است که مجموعه وسیعی از توالي 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.

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

| اصطلاح | تعریف |
|---|---|
| ویرایشگر پایه (Base editor) | ادغام یک پروتئین Cas به یک دامیناز 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 | تناوب‌های کوتاه پالیندرومِ فاصله‌دار منظم خوش‌های، یک منطقه ژنومی باکتریایی که در دفاع از پاتوژن استفاده می شود |
| CRISPRa | استفاده از فعال کننده dCas9 یا dCas9 با gRNA برای افزایش رونویسی یک ژن هدف |
| CRISPRi | استفاده از dCas9 یا سرکوبگر-dCas9 با gRNA برای مانع/کاهش رونویسی یک ژن هدف |
| برش | شکستن دو رشته 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 رشته‌های 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 عموماً 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 می‌شود اما شامل توالی PAM یا توالی gRNA scaffold نمی‌شود. |
| HDR | Homology Directed Repair، یک مکانیسم ترمیم دی‌ان‌ای که از یک الگو برای ترمیم نیک های دی‌ان‌ای یا DSB ها استفاده می‌کند |
| 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. |
| Nick(Nick) | شکست تنها در یک رشته dsDNA |
| Nickase | dsDNA با یکی از دو حوزه نوکلئاز غیرفعال شده است. این آنزیم قادر است تنها یک رشته از هدف را جدا کند. |
| اثرات off-target یا فعالیت off-target | برش Cas9 در مکان‌های 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؛ کدون‌های ترجمه شده که یک ژن را می‌سازند |
| PAM | Protospacer Adjacent Motif؛ توالی مجاور توالی هدف که برای اتصال آنزیم‌های Cas به دی‌ان‌ای هدف ضروری است |
| PCR | Polymerase Chain Reaction؛ برای تقویت و خوانا شدن یک توالی خاص از دی‌ان‌ای استفاده می‌شود |
| 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

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

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

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 می‌کنیم. دسته اول شامل روش‌های مستقیم است که در آن‌ها دانشمندان به رابطه‌های مستقیم بین مکانیزم‌های مختلف و تاثیر آنها روی دقت و حساسیت طرح‌ها، مورد بررسی قرار داده‌اند. دسته دوم روش‌های یادگیری ژرف می‌باشد که برای پیش‌بینی تاثیر و حساسیت طرح‌ها مورد استفاده قرار می‌گیرند.

1.2 روش‌های مستقیم

1.1.2 [34, 3, 2] Chopchop

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 می‌گیرد هم چنین مورد استفاده‌ی طرح را می‌پرسد. به عنوان خروجی لیست مرتب شده طرح‌های ممکن را به همراه offtargets های آن را به ما پس می‌دهد. برای پیدا کردن offtarget از الگوریتمی به نام bowtie استفاده می‌کند و از primer3 برای پیدا کردن primer ها استفاده می‌کند، این الگوریتم با توجه به پژوهش‌های قبلی، از ۶ ویژگی مهم برای مرتب کردن طرح‌ها استفاده می‌کند که عبارت اند از: تعداد offtarget ها، معماری ژن، GC-Content، وجود نوکلوتید G در ۲۰ امین نقطه طرح و همین طور مکان هدف در ژن.

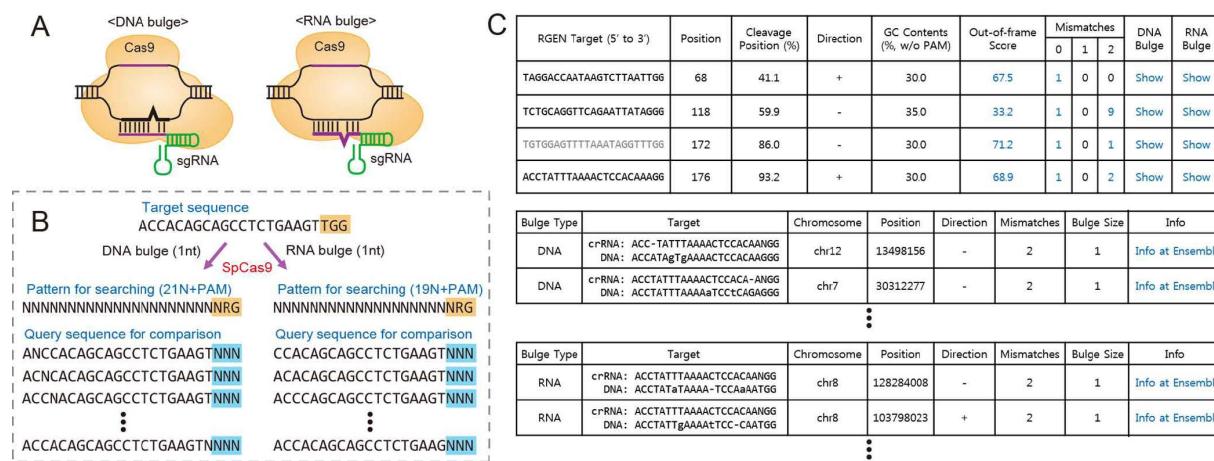
در نسخه دو این الگوریتم، خروجی روی UCSC هم دیده می‌شود و در مورد PAM استفاده شده در هر طرح کاربر اختیار بیشتری دارد و می‌تواند از طرح‌های مختلف CAS استفاده کند. در این نسخه الگوریتم مرتب سازی بر حسب حساسیت و تاثیر طرح‌ها است.

در زمان بین نسخه یک و دو الگوریتم، دانشمندان به نتایج بیشتری رسیده‌اند و از برخی از این نتایج در نسخه جدید الگوریتم chopchop استفاده شده است، این نتایج عبارت اند از: 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 بحسب مکمل بودن دو دویی نوکلوتیدها (نوکلوتید اول دنباله، مکمل نوکلوتید آخر طرح است) است. و در انتهای این امتیاز‌های جدید را با SVM و متريک‌ها مختلف برای مرتب سازی طرح استفاده کردن و اسم آن را امتیاز تاثیرگذاری قرار دادند. 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 آنها امتیاز حساسیت می‌دهد.

2.1.2 [48, 31] Cas-Designer و Cas-OFFinder

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 از CRISPR/Cas نوع II است، که در حال حاضر به طور گسترده برای تحقیقات زیست‌پژوهی و بیوتکنولوژی استفاده می‌شود. Cas-Designer به سرعت فهرستی از تمام توالی‌های آران‌ای راهنمای ممکن در یک توالی دی‌ان‌ای ورودی داده شده ارائه می‌دهد و آنها را در ژنوم انتخابی مشخص می‌کند. علاوه بر این، برنامه امتیاز خارج از چارچوب را به هر نقطه از هدف اختصاص می‌دهد تا به کاربران کمک کند مناطق مناسب برای Knockout ژن انتخاب کنند. Cas-Designer نتایج را در یک جدول تعاملی نشان می‌دهد.

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

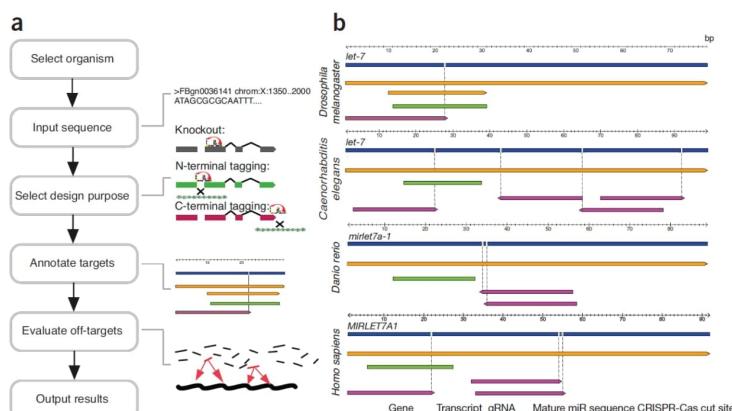


شکل 1.2: (الف) شماتیک مکان‌های off-tagsets را با برآمدگی دی‌ان‌ای یا آران‌ای نشان می‌دهد. (ب) استراتژی برای برآمدگی 1-nt یا آران‌ای بر اساس Cas-OFFinder. (ج) یک مثال از یک جدول خروجی gRNA تمام Cas-Designer ورودی به همراه اطلاعات مفید (بالا) نشان می‌دهد. اگر کاربر روی رنگ آبی کلیک کند عدد، کلمه یا عبارت، اطلاعات دقیق تری مانند اهداف برآمدگی دی‌ان‌ای (وسط) یا آران‌ای (پایین) ارائه می‌شود. علاوه بر این، کاربر می‌تواند موارد مربوطه را به دست آورد اطلاعات ژنومی از طریق مرورگر ژنوم Flicek و همکاران، 2011)، با کلیک بر روی دکمه "اطلاعات در Ensembl [48]".

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

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

[24] E-CRISP 3.1.2

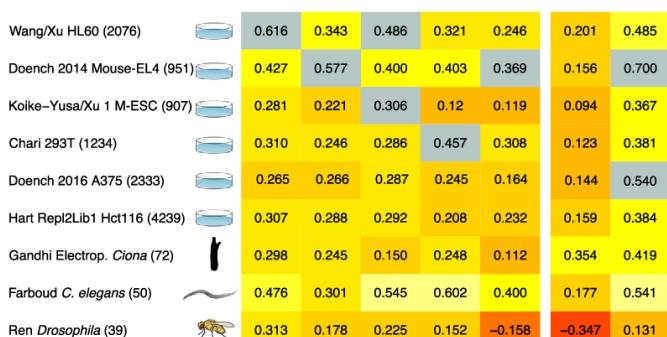


شکل 2.2: الگوریتم E-CRISP [24]

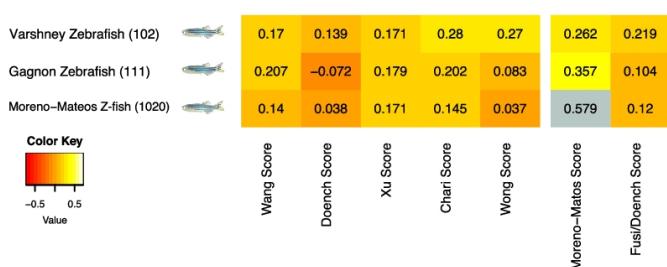
E-CRISP توالی‌های هدف مکمل gRNA را شناسایی می‌کند که به یک موتیف که از سمت ۳' مجاور به G(A) یا N(G) ختم می‌شود، که برای هسته Cas9 مورد نیاز است تا رشتہ دوگانه دی‌ان‌ای را برش دهد. E-CRISP از یک رویکرد نمایه سازی سریع برای یافتن مکان‌های اتصال و یک درخت فاصله دودویی برای حاشیه نویسی سریع سایت‌های هدف gRNA احتمالی استفاده می‌کند. با استفاده از این الگوریتم‌ها، می‌توان در چند ساعت کتابخانه‌هایی در مقیاس ژنومی برای چندین موجود زنده ایجاد کرد.

[30] CRISPOR 4.1.2

Guides transcribed in cells from a U6 promoter



Guides transcribed *in vitro* from a T7 promoter



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

[30]

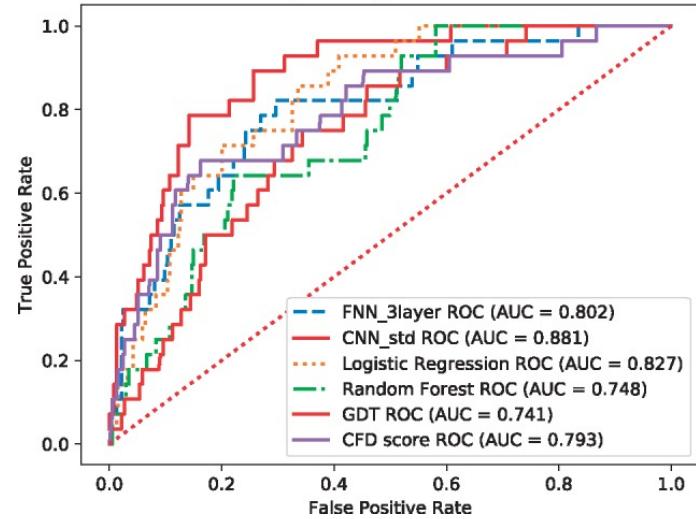
وبسایتی است که به انتخاب و بیان توالی‌های راهنمای CRISPOR کمک می‌کند، که در دو مقاله توضیح داده شده است (CRISPR NAR 2018 و Gen Biol 2016). در حالت پیش‌فرض، کاربر یک توالی دی‌ان‌ای ورودی را چسبانده و ژنوم را انتخاب می‌کند و اطلاعات CRISPOR مربوط به آنها را که در پایگاه‌های اطلاعاتی و الگوریتم‌ها یافت می‌شود، از جمله انواع ژنوم، امتیازهای پیش‌بینی شده off-targets و هدف، را اضافه می‌کند. برای هر دنباله راهنمای، پرایمرهای مختلفی طراحی شده است، به عنوان مثال، برای تقویت هدف، آران‌ای‌های راهنمای را با رونویسی آزمایشگاهی پس از بازپخت پرایمدهای AddGene تولید می‌کند. برای پیش‌بینی، داده‌ها را از هشت مطالعه SpCas9، off-target جمع‌آوری کردن و آنها را با سایت‌های پیش‌بینی شده توسط الگوریتم‌های محبوب مقایسه کردن 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 ۱۰٪ 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 یافته‌ند که امتیاز موثر بودن به شدت به این بستگی دارد که آیا RNA راهنمای 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.

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. روش‌های پیش‌بینی مختلفی توسعه یافته‌اند. با این حال، اکثر آنها فقط امتیازات را بر اساس عدم تطابق با دنباله راهنمای 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 دقیق کافی را در پیش‌بینی‌های خارج از هدف برای ویرایش ژن در سطح بالینی برآورده کنند. برای رفع این مشکل، در این پژوهش دو الگوریتم دو الگوریتم را با استفاده از شبکه‌های عصبی عمیق برای پیش‌بینی جهش‌های off-target در ویرایش ژن CRISPR-Cas9 طراحی و پیاده‌سازی می‌کنیم (با توجه به اطلاعات اولین الگوریتم ماشینی). این مدل‌ها بر روی مجموعه داده‌های اخیراً منتشر شده، مجموعه داده‌های CRISPOR، برای معیار عملکرد، آموخته دیده و آزمایش شدند. یکی دیگر از مجموعه داده شناسایی شده توسط GUIDE-seq به دست می‌آورد، و سطح طبقه‌بندی متوجه زیر نشان می‌دهد که شبکه عصبی کانولوشن بهترین عملکرد دیده توسط داده‌های CRISPOR به دست می‌آورد، و سطح طبقه‌بندی متوجه زیر منحنی ۹۷٪ درصد را تحت اعتبارسنجی متقطع ۵ برابری طبقه‌بندی شده به دست می‌آورد. جالب اینجاست که شبکه عصبی عمیق نیز می‌تواند با میانگین ۹۷٪ در همان تنظیمات رقابتی باشد. ما دو مدل شبکه عصبی عمیق را با روش‌های پیشرفته پیش‌بینی (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

Receiver operating characteristic curve



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

[52]

شده به دست می‌آورد. جالب اینجاست که شبکه عصبی پیشخور عمیق نیز می‌تواند با میانگین ۹۷٪ در همان تنظیمات رقابتی باشد. ما دو مدل شبکه عصبی عمیق را با روش‌های پیشرفته پیش‌بینی (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 لبه‌های رقابتی الگوریتم‌های پیشنهادی است. تحلیل‌های اضافی برای بررسی دلایل زمینه‌ای از دیدگاه‌های مختلف انجام می‌شود.

[49] CCTop 2.2.2

این روش طرح‌های مختلفی که به صورت N20NGG هستند را دسته‌بندی کند، ابتدا با آزمایش‌های عملی طرح‌ها را به دو کلاس موثر و ناموثر دسته‌بندی کرده‌اند. آزمایش به این گونه بود که طرح را در محیط آزمایشگاهی به ژن تزریق می‌کردند و برای هر طرح، تعداد هدف‌های تغییر کرده در طول زمان یادداشت کرده‌اند. این روش بر این باور بود که non-ribosomal و ribosomal RNA بودن ژن در تاثیر طرح موثر است پس دیتابست خود را به دو قسمت تقسیم کرده و برای طرح هر کدام sRNA موثر و ناموثر را تعیین کرده است. این طرح جایگاه هر نیکلوتید را در sRNA های موثر و ناموثر بررسی کرده و به نتایج زیر رسیده است.

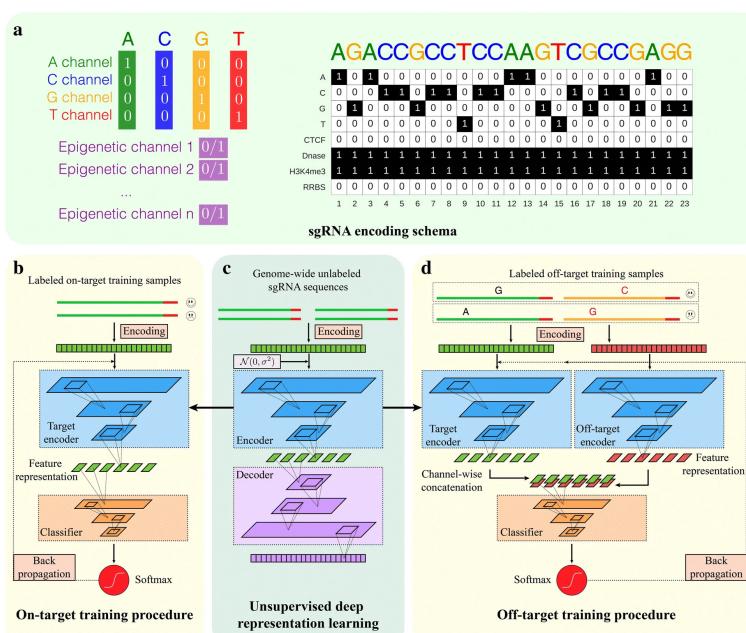
نحوی انتخاب موثر یا ناموثر بودن یک طرح با کمک مدل حسب مدل Elastic-Net است که در آن اگر X_i شده طرح‌ها باشند و \hat{Y}_i امتیاز آن‌ها باشد داریم:

پس از آموختن این مدل، به کمک آزمایش‌های آماری به ۲۸ ویژگی مختلف تاثیرگذاری رسیدند که بیشتر این ویژگی‌ها در ناحیه اسپیسر واقع شده‌اند و بعضی از آنها قبل پیدا شده بود و بعضی جدید بود، که چند نمونه از این دست یافته‌های جدید عبارت اند از:

- قرار گرفتن نیکلوتید G در موقعیت‌های ۱-۲-۳ نسبت به PAM در CAS9 باعث افزایش تاثیرگذاری می‌شود.
- قرار گرفتن نیکلوتید T در چهار موقعیت نزدیک به PAM باعث کاهش تاثیرگذاری می‌شود.
- نوکلئوتیدهای رشتہ ۵' به ۳' تاثیرگذار هستند، در حالی که رشتہ مکمل تاثیر قابل توجهی ندارد.
- قرار گرفتن نیکلوتید C در موقعیت ۳-۴ در CAS9 باعث افزایش تاثیرگذاری می‌شود.
- قرار گرفتن نیکلوتید A در موقعیت ۵-۶ تا ۱۲ باعث افزایش تاثیرگذاری می‌شود.
- قرار گرفتن نیکلوتید G در موقعیت‌های ۱۴ تا ۱۷ باعث افزایش تاثیرگذاری می‌شود.

DeepCRISPR 3.2

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 پیشرفت‌های ترین ابزارهای موجود در سیلیکون است. [26] علاوه بر ویژگی‌های توالی دی‌ان‌ای، چهار ویژگی اپی‌ژنتیکی را معرفی کرد و به طور خودکار اطلاعات معتبر را با استفاده از اصل Auto-encoder استخراج می‌کند. چندین مدل از جمله برش هدف sgRNA و پیش‌بینی تمایل خارج از هدف ایجاد شد. 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 آموختن دادن با در نظر گرفتن اطلاعات اپی‌ژنتیکی و بدون دانستن اطلاعات اپی‌ژنتیکی که نشان می‌دهد که اطلاعات اپی‌ژنتیکی بی‌تأثیر نیست.



[26] DeepCRISPR :5.2

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 |

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

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:

NNNNNNNNNNNNNNNNNNNNNNNNNN

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

NNNNNNNNNNNNNNNNNNNGG

به این رشته ۲۰ تایی، ترکیب ۲۳ تایی PAM می‌گویند که برای sgRNA و ۳ تایی cas9 می‌گویند که برای NGG باشد. تمام کارهای پیشین به روشنی این رشته ۲۳ تایی 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 می‌دهند. 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 می‌باشد که یک مشکل برای ادغام داده‌ها می‌باشد از جمله دو مورد از نتایج نظرات شده هنگام این آزمایش‌ها، شمردن ایندل‌ها 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 با تعداد جایگشت، با تکرار ۴ شی در ۲۱ خانه یا همان $10^{12} \times 4,3980465 \approx 4^{21}$ است. در نتیجه با دیدن حدود ۱۰ هزار یا حتی ۱۰۰ هزار نمونه نمی‌توان نتیجه‌ی عمومی درباره این مسئله گرفت. برای درست کردن روشنی عمومی، به داده‌های زیاد و دقیق نیاز است که در حال حاضر در دسترس نیستند و همین‌طور معمولاً 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. 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 transformer و attention ha برای بدست آوردن یک کدگذاری مناسب است.

از آنجا که بیشتر کارهای پیشین ویژگی‌های دیگر مورد نیاز خود را از ورودی متدانه 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 تنها‌ی کافی نیست، در نتیجه برای رفع این مشکل، ما از روشنی نوین که ایده‌ای مشابه به Stacked Generalization [55] دارد استفاده می‌کنیم تا با مجموعه دادگان کم، دقت بهتری بدست آوریم. می‌توان به روش بدست آمده مانند اصلاح اشتباہات 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 خطاهای را استفاده کردیم و مدل را با آن فاینتون کردیم و با رائی‌گیری بین همه خطاهای مختلف نتیجه‌ی پیشینی مدل را به عنوان بهترین جواب مدل در نظر گرفتیم. با بدست آمدن بهترین نمونه از هر مدل، مدل‌ها را با هم ادغام می‌کنیم تا با اصلاح یک دیگر بهترین دقت را به ما ارائه دهنند. 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.

Learning Ensemble 1.3

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 به تنهایی به دست آورند. [47, 42, 40] بر خلاف آماری، که معمولاً از بی‌نهایت مکانیک آماری استفاده می‌کند، یک مجموعه یادگیری ماشینی تنها از مجموعه محدود مشخصی از مدل‌های تشکیل شده است، اما معمولاً ساختار بسیار انعطاف‌پذیرتری را در بین آن گزینه‌ها امکان می‌دهد.

1.1.3 تعریف

الگوریتم‌های یادگیری نظارت شده وظیفه جستجو در فضای فرضیه را برای یافتن یک فرضیه مناسب انجام می‌دهند که پیش‌بینی‌های خوبی را با یک مسئله خاص انجام دهد. [27]

از زیبایی پیش‌بینی یک مجموعه معمولاً به محاسبات بیشتری نسبت به ارزیابی پیش‌بینی یک مدل نیاز دارد. از بک جهت، یادگیری گروهی ممکن است به عنوان راهی برای جبران الگوریتم‌های یادگیری ضعیف با انجام محاسبات زیاد در نظر گرفته شود. از سوی دیگر، جایگزین این است که یادگیری بسیار بیشتری را در یک سیستم غیر گروهی انجام دهید. یک سیستم ensemble ممکن است در بهبود دقت کلی 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 (مثلًا جنگل‌های تصادفی) استفاده می‌شوند، اگرچه الگوریتم‌های کندرت 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، مقاله‌ای DeepCRISPR دقیقاً داده ۴۲۵ دنباله sgRNA از امتیاز دهنده‌های ۵ مقاله و امتیاز مقاله خود تهیه کرده که از آنها استفاده کردیم. چندین روش ensemble برای جمع این امتیازها و رتبه‌بندی‌ها استفاده کردیم، مانند وزن دهی بر حسب دقت هر مدل روی یک دیتا ثابت و همین طور روش LPA یا Latent Profie Analysis که به ما مدلی برحسب پیش‌بینی مدل‌های دیگر می‌دهند. از این روش‌ها ما دو مدل بدست آوردیم ولی دقت این مدل‌ها همگی از مدل DeepCRISPR پایین‌تر بودند با آنالیز بیشتر به این نتیجه رسیدیم که این مدل‌ها بر سر بعضی نقاط شدیداً اختلاف نظر دارند که باعث تاثیر منفی در نتیجه ensemble این مدل‌ها می‌شود و با این گونه وزن دهی نمی‌توان به نتیجه بهتری رسید. در 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 بررسیم ولی با انجام cross-validation به این نتیجه رسیدیم که دیتای استفاده شده برای آموزش تاثیر زیادی روی دقت پیش‌بینی دارد و لزوماً این روش همیشه از روش DeepCRISPR بهتر نیست، برای بدست آوردن مدل قوی نیاز به دیتای بیشتر داشتیم.

در مرحله‌ی آخر، با توجه به اینکه اکسپرت‌ها اختلاف نظر آنها را کم می‌کرد، چهار الگوریتم، رگرسیون با جنگل تصادفی (Random Forest)، درختان بسیار تصادفی (Extra Trees)، حداقل مربعات معمولی (Ordinary Least Squares)، تقویت گرادیان (Gradient Boosting) را برای ensemble اکسپرت‌ها انتخاب کردیم. هر کدام از الگوریتم‌ها به تنهایی به داده آموزش حساس بودند و با انجام cross-validation لزوماً به نتیجه بهتری نمی‌رسیدند ولی برخلاف اکسپرت‌های اولیه اختلاف نظر این رگرسورها خیلی کم بود پس این متدتها را با هم ادغام کردیم و به نتیجه‌ی مطلوب رسیدیم، یعنی مدلی آموزش داده شده به نویز داده آموزش 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 E-Crisp، CCTop، Cas-Designer، CRISPOR و Chopchop جمع آوری کردیم، که منجر بدست آمدن ۵۰ هزار sgRNA یکتا شد و از آنجا که خروجی الگوریتم‌ها می‌توانست NaN هم باشد، با حذف این داده‌ها به ۳۶ هزار sgRNA یکتا و نظر اکسپرت‌ها راجع به آن رسیدیم. تنها کافی بود که بتوانیم یک standard golden برای این داده‌ها پیدا کنیم، که متساقن‌های قادر به این کار نشیدیم.

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 می‌کند.

مقدمات: آموزش درخت تصمیم

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

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

عمیقی که از قسمت‌های مختلف داده آموزشی ایجاد شده باشند. در این روش معمولاً 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، $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ و $B = (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 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 1 : $b =$

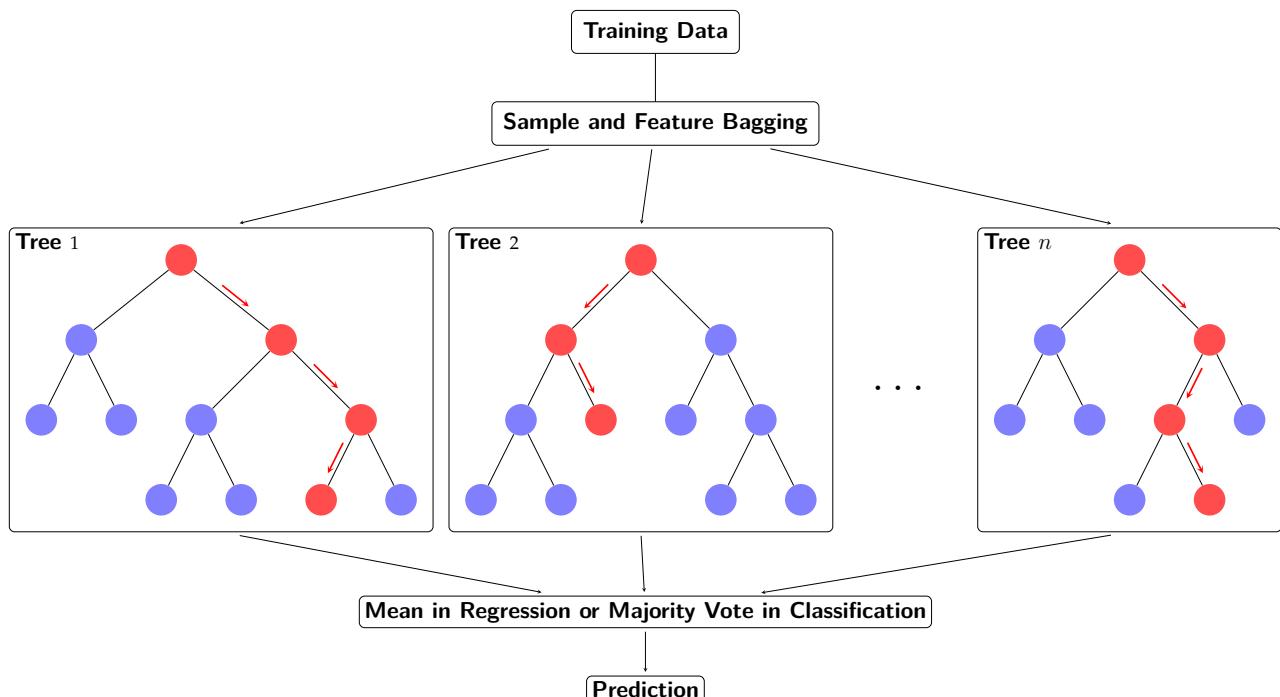
- n نمونه با جایگزینی از داده D انتخاب می‌کنیم و این نمونه‌ها را در مجموعه داده b قرار می‌دهیم. از آنجا که نمونه‌گیری با جایگزینی صورت می‌گیرد یک نمونه ممکن است چندین بار انتخاب شود.

- 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 کاملاً به صورت تصادفی انتخاب می‌کنیم (مثلاً m متغیر اول به مسئله داده شده است، 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، 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 به طور تصادفی برای هر متغیر کاندید شده ترسیم می‌شود و بهترین این آستانه‌های تصادفی تولید شده به عنوان آستانه تقسیم انتخاب می‌شوند. این امر عموماً 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) 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 $\{x_i, y_i\}_{i=1}^n$ داریم. هر مشاهده i شامل یک پاسخ اسکالر y_i و یک بردار ستونی x_i از پارامترهای p (رگرسور)، به عنوان مثال، $x_i = [x_{i1}, x_{i2}, \dots, x_{ip}]^\top$. در یک مدل رگرسیون خطی، متغیر پاسخ، y_i ، یکتابع خطی از رگرسورها است:

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

که از به عنوان بردار به آن نگاه کنیم داریم:

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

به طوری که x_i بردار ستونی از i -امین مشاهده همه متغیرهای است و $\boldsymbol{\beta}$ یک بردار $1 \times p$ 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 می‌توان به صورت نماد ماتریسی نیز نوشت:

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

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

رگرسورها لازم نیست مستقل باشند: هر رابطه دلخواه بین رگرسیون‌ها می‌تواند 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 محدود رگرسیون دیگر است را در نظر می‌گیریم. در آن صورت، مدل در رگرسور 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 قابل محاسبه نیست برای استفاده از این روش معادله زیر را درنظر بگیرید:

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

چنین دستگاهی معمولاً راه حلی برای رسیدن به جواب دقیق ندارد، بنابراین هدف یافتن ضرایب $\boldsymbol{\beta}$ است که نزدیکترین حالت به جواب می‌باشد، 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:

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

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

$$(\mathbf{X}^\top \mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}.$$

به عبارت دیگر:

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

یا

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

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 بهترین وجه متغیر خروجی را از مقادیر متغیرهای ورودی

تقریب می‌کند. این امر با معرفیتابع ضرر $L(y, F(x))$ و به حداقل رساندن آن رسمیت می‌یابد:

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

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) از برحی از کلاس‌های \mathcal{H} ، که یادگیرنده‌گان پایه (یا ضعیف) نامیده می‌شوند:

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

عموماً یک مجموعه آموزشی به ما داده می‌شود $\{(x_1, y_1), \dots, (x_n, y_n)\}$ 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 می‌رساند. این کار را با شروع با یک مدل، مشکل 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.

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

ایده این است که شبیدارترین مرحله فرود را برای این مشکل کمینه‌سازی (نزول شبیب 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 حداقل نزول محلی تابع تلفات، گرادیان منفی است.^[10]

بنابراین، مقدار کمی γ را جایه‌جا 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):

6.1.3 روش پیشنهادی

فرض کنید که برای تخمین رگرسیون از N روش f_1, f_2, \dots, f_N استفاده می‌کنیم. تابع f_i به ازای بردار وزن w و بردار ویژگی x تابع F را تخمین می‌زنند. تابع هزینه را با علامت J نشان می‌دهیم که وابسته به f و w است. در این صورت مسئله هر یک از تخمین‌های رگرسیون f_1, f_2, \dots, f_N برابر است با پیدا کردن بهترین وزن‌ها نسبت به تابع هزینه یعنی:

$$\arg \min_w \sum_{j=1}^n J_{f_i, w}(x_j, y_j)$$

که در آن (x_j, y_j) داده‌های آموزش ما هستند. حال یک پیچیدگی دیگر نیز به مسئله اضافه می‌کنیم، فرض کنید M تابع هزینه داریم و می‌خواهیم از تمام روش‌ها و توابع هزینه بهترین استفاده را ببریم. ابتدا برای ساده سازی روش f_i را به دو قسمت تقسیم می‌کنیم،

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:

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

- 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:

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

- Update the model:

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

end

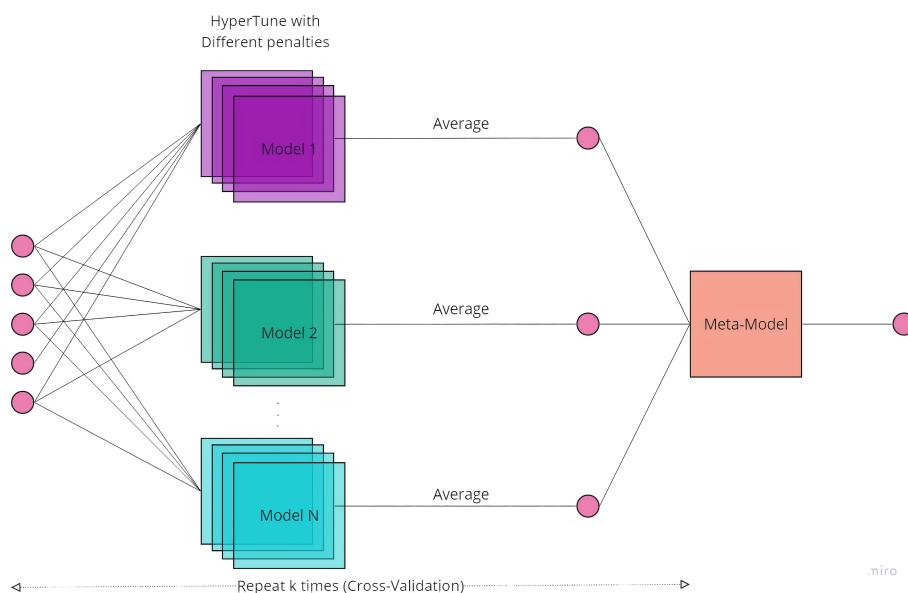
قسمتی که تنها به x_i وابسته است و قسمتی دیگر، در نتیجه داریم:

$$f_i(x; w) = g_i(x) + l_i(x, w)$$

در نتیجه با تغییر تابع J فقط تابع l_i تغییر می‌کند و یکی از ابتدایی ترین روش‌ها برای استفاده از تمام توابع هزینه مختلف میانگین گرفتن تمام جواب‌های بدست آمده از این توابع هزینه است:

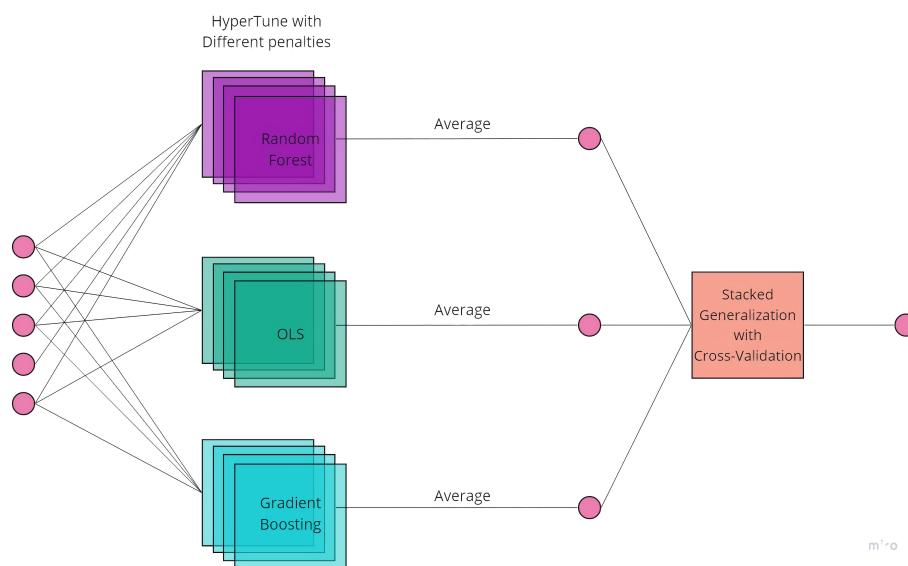
$$\begin{aligned} \operatorname{avg}_{J_j} f_i(x; w^{(J_j)}) &= \frac{1}{M} \sum_{j=1}^M g_i(x) + l_i(x, w^{(J_j)}) = g_i(x) + \frac{1}{M} \sum_{j=1}^M l_i(x, w^{(J_j)}) \\ &= g_i(x) + l_i^{(J^*)}(x, w^{(J^*)}) = f_i(x; w^{(J^*)}) \end{aligned}$$

یا به عبارتی دیگر، اگر f_i خطی باشد، یک تابع هزینه جدید برای روش f_i محاسبه می‌شود که سعی می‌کند میانگین توابع هزینه تعیین شده راه در حالت کمینه نگه دارد. در نظر داشته باشید که در این راه ما انتخاب کردیم که از روشی خطی برای ادغام توابع هزینه استفاده کنیم چون شهود بیشتری نسبت به آن داشتیم ولی در ادامه می‌بنیم که با اینکه این کار حتی به روش‌های غیر خطی نیز کمک می‌کند و می‌توانیم از هر روشی برای ادغام به اصطلاح توابع هزینه استفاده کنیم. حال با بدست آمدن یک پیش‌بینی از هر f_i ، با ادغام آنها با هم می‌توانیم یک پیش‌بینی از با کمک همه توابع هزینه و همه روش‌ها داشته باشیم. فقط دقت داشته باشید که از تمام روش‌های ادغام استفاده شده خطی باشند در این صورت روش به ادغام همه روش‌ها با هم خلاصه می‌شود و به همین دلایل پیشنهاد می‌شود که ادغام روش‌های f_i را به صورت غیرخطی انجام دهید.



شکل 1.3: شمای روش پیشنهادی

با توجه به روش ارائه شده ما مدلی بدین شکل درست کردہ‌ایم:



شکل 2.3: شمای دقیق استفاده شده برای بدست آوردن نتیجه

7.1.3 آنالیز مشخصات پنهان

آنالیز مشخصات پنهان (به انگلیسی: Latent Profile Analysis) (LPA)، یک رویکرد مدل‌سازی آماری برای تخمین پروفایل‌های متمایز متغیرها است. در علوم اجتماعی و در تحقیقات آموزشی، این پروفایل‌ها می‌توانند به عنوان مثال نشان دهنده که چگونه سن‌های مختلف در آزمایش تاثیرگذار بوده اند. توجه داشته باشید که LPA با متغیرهای پیوسته (و در برخی موارد، متغیرهای ترتیبی) کار را دارد، اما برای متغیرهای دوگانه (دودویی) مناسب نیست. ما به کمک برنامه‌ی tidyLPA شیش مدل مختلف را تست کردیم، همانطور که قبل اهل اعلام کردیم، فرضیات LPA را در نظر نداشته‌ایم و فقط به صورت آزمایشی آن را امتحان کردیم. به طور کلی، رویکرد انتخاب مدل مشابه انتخاب تعداد پروفایل‌ها است، که مستلزم تصمیم‌گیری بر اساس شواهد از منابع متعدد، از جمله معیارهای اطلاعاتی، آزمون‌های آماری، و تفسیرپذیری است.

در tidyLPA، شش مدلی که امکان تعیین آنها در LPA وجود دارد از نظر چگونگی تخمین متغیرهای مورد استفاده برای ایجاد پروفایل‌ها توضیح داده شده است.

۱. واریانس پروفایل‌ها برابر و کواریانس آنها برابر با صفر است (مدل ۱) ۲. واریانس پروفایل‌ها متفاوت و کواریانس آنها برابر با صفر است (مدل ۲) ۳. واریانس پروفایل‌ها برابر و کواریانس آنها نیز برابر است (مدل ۳) ۴. میانگین پروفایل‌ها متفاوت، واریانس آنها متفاوت و کواریانس آنها برابر است (مدل ۴) ۵. میانگین پروفایل‌ها متفاوت، واریانس آنها برابر و کواریانس آنها متفاوت است (مدل ۵) ۶. واریانس پروفایل‌ها متفاوت و کواریانس آنها متفاوت است (مدل ۶)

هنگام استفاده از پکیج tidyLPA ، کافی است مدل خود و تعداد پروفایل‌ها را انتخاب کنید و پکیج در صورت همگرا بودن جواب، آن را نمایش می‌دهد، ما تمام ۶ مدل را برای دو پروفایل‌های sgRNA موثر و ناموثر امتحان کردیم که فقط دو مدل جواب همگرا داشتند و آن‌ها هم مدل ۱ و ۲ بودند.

Attention 2.3

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

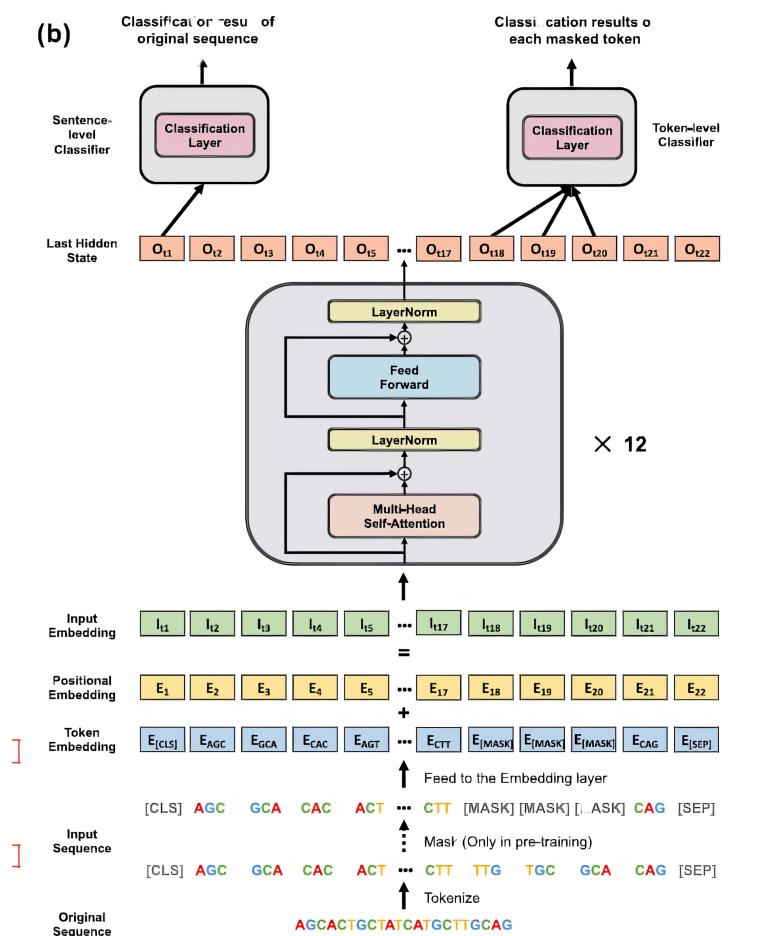
در شبکه‌های عصبی، توجه تکنیکی است که توجه شناختی را تقلید می‌کند. این اثر باعث می‌شود که اثر برخی از بخش‌های ورودی افزایش یابد در حالی که بخش‌های دیگر را کاهش می‌دهد - فکر این است که شبکه باید تمرکز بیشتری را به آن بخش کوچک اما مهم داده اختصاص دهد. یادگیری اینکه کدام بخش از داده‌ها مهم تر از سایرین است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

مکانیسم‌های مانند توجه در دهه ۱۹۹۰ با نام‌هایی مانند ماذول‌های ضربی، واحدهای سیگما پی و ابرشبکه‌ها معرفی شدند. [36] انعطاف‌پذیری آن ناشی از نقش آن به عنوان "وزن نرم" است که می‌تواند در طول زمان اجرا تغییر کند، برخلاف وزنه‌های استاندارد که باید در زمان اجرا ثابت بمانند. کاربردهای توجه شامل حافظه در ماشین‌های تورینگ عصبی، وظایف استدلال در رایانه‌های عصبی متمایز [21]، پردازش زبان در ترانسفورماتورها، و پردازش داده‌های چندحسی (صدا، تصاویر، ویدئو، متن) در درک‌کننده‌ها است. [29, 50, 45, 44]

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 و roberta sgRNA استفاده کردیم که نتایج این مدل‌ها خیلی ضعیف بود. با توجه به آنالیزهای انجام شده به این نتیجه رسیدیم که مشکل از دیتاها بودن برچسب 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 کلمات تایی k را به عنوان دیکشنری کلمات در نظر می‌گیریم. برای قسمت pretrain از sgRNA که خودمان ذخیره کرده بودیم و داده‌های دیگر استفاده کردیم و سپس برای تنظیمات نهایی از داده‌های مقاله DNAbert استفاده کردیم ولی نتایج آن نتایج جالبی نبود.

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 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.



[60] شکل 3.3: مدل دقت DNAbert

فصل 4

نتایج شبیه‌سازی

ابتدا نمایشی از داده‌های گلدن استاندارد و تفاوت آنها در آزمایش‌ها مختلف را نشان می‌دهیم:

جدول 1.4: تفاوت اندازه‌گیری فرکانس ایندل‌ها برای یک سل لاین در دو پژوهش مختلف

| gRNA+PAM | Wang et al. (2019) Indel_freq_HEK293T | Kim et al. (2019) Indel_freq_HEK293T |
|--------------------------|--|---|
| GAGGAAGCAGATATCCGGTGTGG | 94.313725490196006 | 40.490007577838398 |
| GGAGGAGGCTGAACGCCACGAGGG | 90.129016553067203 | 74.369471837500697 |
| GCTGCGAGACCGCTATCCCGTGG | 94.1153758800817 | 85.4296388542964 |
| GCGCGTCGAACACGAACCAGCGG | 94.067796610169395 | 61.945179048985402 |
| ATACTCACATCACAGCCCCGTGG | 45.164998674709402 | 43.601387998980698 |
| GAATACGCCTCTGCCTTCAAGG | 42.752196781612703 | 24.8797250859107 |
| GACAGTGCACCGTGACGTGG | 86.950586950586896 | 87.678945915304297 |
| GTCCCAACTCCTGCGCACGAAGG | 87.792680154580495 | 78.251019483461704 |
| GTATGTCGAGAGTACCAACGTGG | 93.989694643289397 | 84.282105733435799 |
| GAAGTCCCAGAATGACTCCTGTGG | 95.953478478770094 | 51.447561838907902 |
| GCAAGAGCTCTCAATTACACAGG | 26.400666586386201 | 41.090027521361897 |
| GACCTACCACCGAGCCATCAAGG | 45.699392752721998 | 48.920244981226801 |
| ATTCTTACAGACAGGTCCGGTGG | 71.398959583833502 | 58.898283855940598 |
| ATTCCAGATCCAAGTGCAGAAGG | 19.416422401075099 | 30.0529172782263 |
| ATAACCTGTAAGCCCCACAAAGG | 84.049773755656105 | 72.983725135623899 |
| GAGCATGCCAGCACGCTCAAGGG | 35.8299328682374 | 68.013567829869004 |
| GGAAGCCGAGATCCCCCGCGAGG | 96.926026719445801 | 50.104123281965897 |
| GGTCCCTTAGCTCTCATGTGG | 65.196962505932603 | 60.4208849810732 |
| GGACCGGGAAAGCAATTGACAGG | 60.6481507594943 | 48.141659670510599 |
| AGCGTAAGCCAATACTGATGAGG | 60.069097691358699 | 58.825459317585299 |
| GCTTCCAAGTAGCACTCAGTAGG | 50.230952263469199 | 43.932746018438102 |
| GCTGCACTACTACCCGACGTGG | 88.571428571428498 | 70.507599887110402 |
| AAATCTTGTGAACCTCATCGAGG | 76.194029850746205 | 42.066691095639399 |

در اینجا ما از داده‌های ارائه شده در مقاله DeepCRISPR برای مقایسه مدل‌های مختلف استفاده کرده‌ایم که حدود ۴۳۰ 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.

جدول 2.4: داده‌های مقایسه امتیاز الگوریتم‌های مختلف [26]

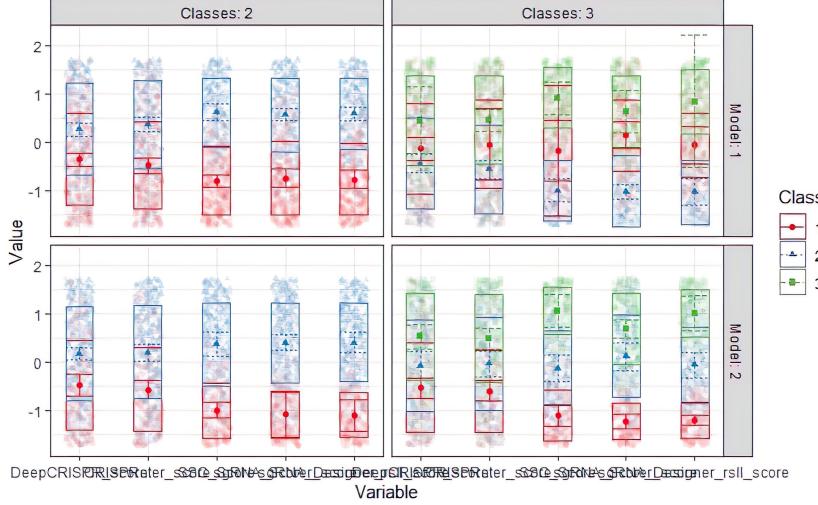
| sgRNA number | KO reporter assay | DeepCRISPR score | CRISPRater score | SSC Score | sgRNA Scorer score | sgRNA Designer rsII score | sgRNA sequence | extended spacer |
|--------------|-------------------|------------------|------------------|-----------|--------------------|---------------------------|----------------------------|--------------------------------|
| sg1 | 0 | 0.17706534 | 0.571 | -0.485 | 30.66 | 0.571 | GAGTCGGGGTTCTCATGTTGG | AGTAGACTGGGGTTCTCATGTTGGCA |
| sg2 | 0 | 0.055156678 | 0.6998 | -0.266 | 54.96 | 0.533 | CGCCGCCGCTTCCGGTATGAGG | CTGCGCCGCCGCTTCCGGTATGAGGAAA |
| sg3 | 0 | 0.23954645 | 0.6865 | -0.448 | 25.79 | 0.41 | GGCACCGGTGTCACGGGTCGGG | CCCGGGCACGGTCTGCACGGGTCGGGTA |
| sg4 | 0 | 0.147778 | 0.6405 | -4.6E-2 | 53.81 | 0.491 | TGGGGGATCACTTGACGTCAAGG | GAGGTGGGGATCACTTGACGTCAAGGAGT |
| sg5 | 0 | 0.121 | 0.68 | 6.7E-2 | 12.44 | 0.485 | TTACCATAGTGTACGGGTGAGG | CCTTTACCATAGTGTACGGGTGAGGCAT |
| sg6 | 0 | 0.14186779 | 0.5489 | 0.085 | 64.75 | 0.489 | TCTACTGAAGTGTAGCAAACAGG | TCTCTACTGAAGTGTAGCAAACAGGTAC |
| sg7 | 0 | 0.10871141 | 0.6207 | 0.107 | 24.01 | 0.554 | TAGAGATCCGCCATCTCAAGG | CAGATAGAGATCCGCCATCTCAAGGGAC |
| sg8 | 0 | 0.14419994 | 0.6916 | 0.91 | 73 | 0.448 | CTCATACCGAAAGGCCGGCG | TTCTCTCATACCGAAAGGCCGGCGCAG |
| sg9 | 0.028 | 0.11949389 | 0.5259 | -0.578 | 9.73 | 0.441 | TTCTGAATTATCGCTAGCCTGG | AGATTTCGAAATTATCGCTAGCCTGGAT |
| sg10 | 0.036 | 0.151749 | 0.4501 | -0.329 | 52.9 | 0.412 | GCCTCAGGCTCAGGAATAGCTGG | GCCTGCCTCAGGCTCAGGAATAGCTGGAT |
| sg11 | 0.037 | 0.13260305 | 0.5663 | -0.364 | 10.57 | 0.521 | AAAGTACTCTGGAGTACTGCAGG | CCCAAAGTACTCTGGAGTACTGCAGGAGG |
| sg12 | 0.056 | 0.28234875 | 0.7611 | 0.019 | 83.17 | 0.624 | CACCGTAGTCATCTCAATGAGG | AGATCACCGTAGTCATCTCAATGAGGGCC |
| sg13 | 0.064 | 0.099274084 | 0.6184 | 0.318 | 15.79 | 0.504 | ACGGAGTCGCTGTCGCCAGG | TGAGACGGAGTCGCTGTCGCCAGGCTG |
| sg14 | 0.066 | 0.14608404 | 0.6311 | 0.002 | 51.42 | 0.395 | TGGGATGCCGTCCCGAAAAATGG | GTAGTGGGATGCCGTCCCGAAAAATGGCC |
| sg15 | 0.072 | 0.098504 | 0.6751 | 0.587 | 79.21 | 0.599 | TCCGAGAGAACCTCGCAAGGG | CAGATCCGAGAGAACCTCGCAAGGGATT |
| sg16 | 0.109 | 0.14631987 | 0.5128 | -0.327 | 55.6 | 0.27399 | CCGTCAGGCCAGCGAACGCTGG | GCCCCCGTCAGGCCAGCGAACGCTGGCT |
| sg17 | 0.111 | 0.19384801 | 0.8201 | -0.249 | 22.99 | 0.505 | CTAGTGGAAAGTGAACGCTCTGG | TGGACTAGTGGAAAGTGAACGCTCTGGCAT |
| sg18 | 0.113 | 0.20544451 | 0.5886 | -0.102 | 43.12 | 0.561 | GGGCATATGGACTAGGCACTGGG | TGTGGGCATATGGACTAGGCACTGGCTA |
| sg19 | 0.125 | 0.097481444 | 0.6024 | -0.278 | 3.89 | 0.438 | TGACATTCAATTCGCTAGCTGG | AACTGACATTCAATTCGCTAGCTGGACA |
| sg20 | 0.137 | 0.24448508 | 0.6321 | 0.374 | 78.21 | 0.661 | GCTTACCACTATGACGAGCATGG | GTGTGCTTACCACTATGACGAGCATGGTA |
| sg21 | 0.14 | 0.17448096 | 0.7384 | 1.44 | 97.25 | 0.718 | GTTAGGAAATCGTCACCCGGCGG | CGCGGTTAGGAAATCGTCACCCGGCGCT |
| sg22 | 0.156 | 0.19983143 | 0.6321 | 0.013 | 45.91 | 0.521 | GCTAACGATCTTTGATGATGG | TCTCTGCTAACGATCTTTGATGATGGCT |
| sg23 | 0.161 | 0.18330452 | 0.6786 | 0.155 | 24.17 | 0.425 | ACCAGTACAAACGGGCTCGG | TCCCACCAGTTACAAACGGGCTCGGCT |
| sg24 | 0.162 | 0.1040944 | 0.4028 | -0.512 | 30.01 | 0.393 | AGCTTACCAAGGCTAGAGTGCCTAGG | AAGCAGCTACCAAGGCTAGAGTGCCTAGG |
| sg25 | 0.179 | 0.16135208 | 0.6374 | -0.241 | 49.36 | 0.215 | TCGGCTGAAATAITGTTIAAGG | TCTATCGCTGAAATAITGTTIAAGGATT |

LPA 1.4

اولین آزمایشی که در روش LPA انجام دادیم این بود که بینینم کدام یک از مدل‌ها همگرا می‌شوند:

جدول 3.4: نتیجه اجرای مدل‌های مختلف بر تقسیم به ۳ یا ۲ پروفایل

| Model | Classes | LogLik | AIC | AWE | BIC | CAIC | CLC | KIC | SABIC | ICL | Entropy | prob_min | prob_max | n_min | n_max | BLRT_val | BLRT_p |
|-------|---------|-----------|----------|----------|----------|----------|----------|----------|----------|-----------|---------|----------|----------|---------|---------|----------|--------|
| 1 | 2 | -2787.420 | 5606.840 | 5814.917 | 5671.673 | 5687.673 | 5576.430 | 5625.840 | 5620.900 | -5730.746 | 0.795 | 0.937 | 0.944 | 0.44471 | 0.55529 | 450.642 | 0 |
| 1 | 3 | -2759.128 | 5562.256 | 5849.152 | 5651.402 | 5673.402 | 5519.652 | 5587.256 | 5581.588 | -5793.690 | 0.698 | 0.795 | 0.895 | 0.28706 | 0.36941 | 56.585 | 0 |
| 5 | 2 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| 5 | 3 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| 3 | 2 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| 3 | 3 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| 2 | 2 | -2777.674 | 5597.348 | 5870.764 | 5682.442 | 5703.442 | 5557.120 | 5621.348 | 5615.801 | -5713.236 | 0.886 | 0.947 | 0.977 | 0.27765 | 0.72235 | 470.136 | 0 |
| 2 | 3 | -2706.045 | 5476.090 | 5893.834 | 5605.757 | 5637.757 | 5413.680 | 5511.090 | 5504.209 | -5701.814 | 0.795 | 0.890 | 0.939 | 0.21176 | 0.50824 | 143.589 | 0 |
| 6 | 2 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| 6 | 3 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| 4 | 2 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| 4 | 3 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |



شکل 1.4: پراکندگی داده نسبت به کلاس‌ها

مقایسه این مدل‌ها با کارهای پیشین نشان داد که روش قابل قبول نیست.

جدول ۴.۴: داده‌های مقایسه امتیاز الگوریتم‌های مختلف [26]

| | Model 1 | *Ours | DeepCrispr | CRISPRater | SSC | sgRNA Scorer | sgRNA Scorer |
|-------------------------|----------|----------|------------|------------|----------|--------------|--------------|
| Thersholt = 0.6 | | | | | | | |
| Accuracy Score: | 0.648140 | 0.914763 | 0.739884 | 0.660604 | 0.586359 | 0.580693 | 0.598542 |
| Precision Recall Score: | 0.715779 | 0.936461 | 0.807515 | 0.714867 | 0.648901 | 0.662849 | 0.673465 |
| F1 Score: | 0.557377 | 0.875740 | 0.0 | 0.541463 | 0.159468 | 0.762044 | 0.078571 |
| Thersholt = 0.7 | | | | | | | |
| Accuracy Score: | 0.631560 | 0.880356 | 0.684054 | 0.642937 | 0.563159 | 0.568745 | 0.623188 |
| Precision Recall Score: | 0.531786 | 0.823824 | 0.581910 | 0.520233 | 0.440453 | 0.476842 | 0.494809 |
| F1 Score: | 0.489297 | 0.712329 | 0.0 | 0.167488 | 0.108911 | 0.579564 | 0.0 |
| Thersholt = 0.8 | | | | | | | |
| Accuracy Score: | 0.521924 | 0.830463 | 0.635765 | 0.608747 | 0.427671 | 0.475937 | 0.585022 |
| Precision Recall Score: | 0.185667 | 0.503301 | 0.168832 | 0.148355 | 0.091240 | 0.107466 | 0.179785 |
| F1 Score: | 0.171123 | 0.156863 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 |

شايان ذكر است که چون در روش LPA از کل داده‌ها برای آموزش استفاده می‌شود، برای حساب کردن امتیاز سایر الگوریتم‌ها نیز از کل داده استفاده کردیم و از آنجا که روش پیشنهادی (Ours) نیز با همین داده‌ها آموزش داده شده پس این مقایسه، مقایسه‌ی خوبی برای بررسی توانایی روش پیشنهادی نیست.

2.4 روش پیشنهادی

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 (کلاس بندی).

حال نتیجه روش پیشنهادی برای ادغام متدهای پیشین که با تقسیم ۷۰ به ۳۰ 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 ۱۰۰ تقسیم تصادفی امتحان کرده‌ایم و میانگین هر کدام را ارائه می‌دهیم:

جدول ۵.۴: مقایسه کامل روش DeepCRISPR با روش‌های مختلف Ensemble

| Regression | | | | | | | | |
|-----------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------------|--------------------------|--------------------------|
| | *OURS | DeepCRISPR | RandomForest | LinearRegression | GradientBoosting | Average RandomForestRegressor | Average LinearRegression | Average GradientBoosting |
| spearman_score | 0.48363014255265202 | 0.44920573337131903 | 0.442334582432172 | 0.48875954923704201 | 0.42820701429086999 | 0.445627937403689 | 0.486785109490944 | 0.43753255749096398 |
| MSE_score | 4.3495596387515302E-2 | 8.8698237964087503E-2 | 4.07988086339467E-2 | 4.4421889409931303E-2 | 4.2615593436146702E-2 | 5.770390892540909E-2 | 4.5011123276089499E-2 | 5.0119844259733599E-2 |
| Classification | | | | | | | | |
| Thersholt: 0.7 | *OURS | DeepCRISPR | RandomForest | LinearRegression | GradientBoosting | Average RandomForestRegressor | Average LinearRegression | Average GradientBoosting |
| accuracy_score | 0.65585937500000002 | 0.53796875 | 0.677734375 | 0.645390625 | 0.69515625000000003 | 0.67257812500000003 | 0.6415625000000004 | 0.61054687500000004 |
| roc_auc_score | 0.66431776673864995 | 0.61734640570578403 | 0.65843505577961503 | 0.65901999368111197 | 0.67276455067103103 | 0.65458082510786497 | 0.65880260125869 | 0.6391789305687402 |
| precision_score | 0.79255770213534 | 0.85072046233693299 | 0.76624039613931905 | 0.79259641935763803 | 0.77357966284998403 | 0.76411257617980399 | 0.79565743804489097 | 0.79178486827439998 |
| recall_score | 0.63734210988092399 | 0.34806419545723799 | 0.72518684127973998 | 0.614432263387238 | 0.74964082726952797 | 0.71661369491892901 | 0.60264418676158005 | 0.55093050984697001 |
| f1_score | 0.703752997266619 | 0.49253449071580302 | 0.74347366186999897 | 0.690397602288365 | 0.76016707914694104 | 0.73783309071755199 | 0.68385488400419403 | 0.64166082655091905 |
| Classification | | | | | | | | |
| Thersholt: 0.8 | *OURS | DeepCRISPR | RandomForest | LinearRegression | GradientBoosting | Average RandomForestRegressor | Average LinearRegression | Average GradientBoosting |
| accuracy_score | 0.62843749999999898 | 0.60109374999999898 | 0.60835937500000004 | 0.63249999999999895 | 0.61617187500000004 | 0.53515625 | 0.63835937499999895 | 0.53507812499999896 |
| roc_auc_score | 0.57763471246611797 | 0.57763471246611797 | 0.59211535894568601 | 0.61618250051450296 | 0.60167042790898995 | 0.50032965589703104 | 0.622627094031456 | 0.5 |
| precision_score | 0.61123631849945304 | 0.70734574295302999 | 0.6407490528634304 | 0.69322291365585 | 0.64869297561177897 | 5.4545454545454948E-3 | 0.6975096155772995 | 0 |
| recall_score | 0.61123631849945304 | 0.24286032851065301 | 0.35517924784142701 | 0.38067809464656399 | 0.38781860176996902 | 1.2765957446808499E-3 | 0.39614240212076302 | 0 |
| f1_score | 0.61123631849945304 | 0.35955398993128102 | 0.45563124591441601 | 0.48885398051706602 | 0.48211167798519799 | 2.0689655172413798E-3 | 0.50302048323802395 | 0 |
| Classification | | | | | | | | |
| Thersholt: 0.9 | *OURS | DeepCRISPR | RandomForest | LinearRegression | GradientBoosting | Average RandomForestRegressor | Average LinearRegression | Average GradientBoosting |
| accuracy_score | 0.80203124999999897 | 0.77585937500000002 | 0.80976562500000004 | 0.80156249999999896 | 0.80390625000000004 | 0.80718749999999895 | 0.79328125000000005 | 0.80718749999999895 |
| roc_auc_score | 0.56430742251440802 | 0.50133606125879804 | 0.51214587901508801 | 0.5702464886899598 | 0.5116932503973103 | 0.5 | 0.59023109127478102 | 0.5 |
| precision_score | 0.46911614774114702 | 0.20470057720057699 | 0.41133333333333 | 0.46759052059051998 | 0.3483055555555498 | 0 | 0.44165478011840198 | 0 |
| recall_score | 0.17702155764350999 | 5.4478758121587097E-2 | 2.7348192951095199E-2 | 0.193655913131485 | 3.573162726006601E-2 | 0 | 0.25916550040003999 | 0 |
| f1_score | 0.24884028248654999 | 8.40075202678989E-2 | 5.0123824028299999E-2 | 0.26732121294272798 | 6.2823477586924303E-2 | 0 | 0.32175545861810201 | 0 |

Attention 1.2.4

برای روش‌هایی که فقط از دنباله sgRNA استفاده می‌کنیم، ابتدا حدود ۴ میلیون sgRNA از دادگان کارهای پیشین و زن‌های مختلف جمع‌آوری کردیم و با ترنسفورمرهای تک حرفی و چندحرفی ای 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 بست‌آمده را بروی دیتا با تقسیم threshold ۸۰ به ۲۰ و ۷٪ کلاس‌بندی کردیم. نتیجه‌ی دسته بندی بعد از آموزش به کمک مدل‌های توجه

| DNAbert Attention Model | | | |
|-------------------------|---------------------|---------------------|---------------------|
| | 3mer | 4mer | 6mer |
| Accuracy | 0.709879518072289 | 0.709879518072289 | 0.709879518072289 |
| AUC | 0.50249169435215901 | 0.503859617071856 | 0.503859617071856 |
| F1 | 0.41516347237880402 | 0.41516347237880402 | 0.41516347237880402 |
| MCC | 0 | 0 | 0 |
| Precision | 0.35493975903609998 | 0.35493975903000002 | 0.354939759036144 |
| Recall | 0.5 | 0.5 | 0.5 |

جدول 4.6: نتیجه‌ی آموزش مدل‌های DNAbert برای کلاس‌بندی های sgRNA موثر و ناموثر

نتیجه آموزش مدل توجه برای بست‌آوردن بردار کد

| Epoch | Training Loss | Validation loss | Accuracy |
|-------|---------------|-----------------|--------------------|
| 1 | 0.665000 | 0.724736 | 0.5619589999999999 |
| 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 | 0.65900 | 0.701253 | 0.561959 |
| 7 | 0.656900 | 0.733162 | 0.561959 |
| 8 | 0.650700 | 0.72118 | 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 | 0.70978 | 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 | 0.6462 | 0.694035 | 0.561959 |
| 18 | 0.655700 | 0.6896 | 0.561959 |
| 19 | 0.645500 | 0.708879 | 0.561959 |
| 20 | 0.646800 | 0.706368 | 0.561959 |

جدول 4.7: نتیجه تمرین به کمک 6mer، به کمک مدل RoBerta

فصل 5

جمع‌بندی و کارهای آتی

دو مشکل اساسی که در داده‌ها پیدا می‌شود: ۱) نویز ذاتی داده‌ها که به خاطر حضور یک sgRNA در cell-line ها و ارگانیزم‌ها مختلف است، چون که یک ارگانیزم خاص می‌تواند خیلی خوب عمل کند ولی در ارگانیزم دیگر عمل کرد متوسط و یا ضعیفی داشته باشد. ۲) نامتعادل بودن داده‌ها است به خاطر بایاس پژوهشگران هنگام انتخاب sgRNA یا نحوی اندازه‌گیری امتیاز تاثیرگذاری نیز یک مشکل بزرگ برای پیش‌بینی دقیق به حساب می‌آید. مثلاً معمولاً کارشناسانی که sgRNA های مختلف را تست می‌کنند، یک حس از قبل روی میزان تاثیرگذاری این ها sgRNA دارند و به همین دلیل ای sgRNA که فکر می‌کنند اصلاً خوب عمل نمی‌کنند را آزمایش نمی‌کنند که باعث به وجود آمدن نامتعادل دیتاست‌های نامتعادل می‌شود.

از جمله کارهایی که می‌توان برای حل این مشکلات انجام داد این است که روش پیشنهادی را به جای اینکه با ورودی متدهای دیگر پیاده‌سازی کنیم، روی ویژگی‌های بدست آمده پیاده سازی شود و مستقیماً سعی به بهبود رگرسیون کنیم و یا با ادغام مدل‌هایی که روی ارگان‌های مختلف آموزش داده شدند، مدلی عمومی‌تر درست کرد. البته برای این کار، نیاز به داده‌ی زیادی است که تمام ویژگی‌های مختلف پیدا شده را پوشش دهد، با توجه به تجربه، عموماً برای هر یک ویژگی حداقل ۱۰۰ نقطه نیاز است تا نتیجه مطلوب باشد. علاوه بر آن به نظر می‌رسد که تعداد ویژگی‌های پیدا شده بسیار بالا‌س. در نتیجه، باید بدنیال روشی برای انتخاب ویژگی‌های بهینه هم باشیم. یک روش جالب برای حذف نویز، دیدن sgRNA مانند یک تصویر است. همانطور که کدگذاری One-Hot برای پیدا کردن امتیاز بهتر در دنباله‌های خارج از هدف استفاده شده است، می‌توان به دنبال روش‌های جالب‌تر و پیچیده‌تری برای نمایش sgRNA و چند ویژگی دیگر به صورت تصویر اشاره کرد، تا بتوان به کمک روش‌های یادگیری ژرف، پ شبیه‌ی بهتری انجام داد. البته عموماً روش‌های یادگیری ژرف نیز به تعداد داده‌ی بالایی نیاز دارند. از نمونه روش‌هایی که می‌توان طول دنباله‌ی sgRNA را هنگام تبدیل به تصویر، بلندتر کرد، روش kmer و کدکردن ارگان و cell-line همراه با sgRNA است.

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-Barwińska, 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_lets_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 Concorde, 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, Wen-hui 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).
- [60] Yanrong Ji, Zhihan Zhou, Han Liu, Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. doi:10.1093/bioinformatics/btab083 (2021).
- [61] Wang, D. et al. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nat. Commun.* 10, 4284 (2019).
- [62] Kim, N. et al. Prediction of the sequence-specific cleavage activity of Cas9 variants. *Nat. Biotechnol.* 38, 1328–1336 (2020).

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 method, 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

April 6, 2022