

فصل ۱

مقدمه

تومور^۱ از رشد غیر طبیعی سلول با احتمال حمله یا گسترش به سایر قسمت‌های بدن تشکیل می‌شود. تومورهای بدخیم^۲ معمولاً سرطان^۳ نامیده می‌شوند. سرطان علل مختلفی از جمله تغییرات ژنتیکی، آلودگی محیط زیست یا انتخاب‌های نادرست در سبک زندگی دارد. یک تومور ممکن است از زیرجمعیت‌های سلولی با تغییرات ژنومی مشخص تشکیل شده باشد، این پدیده ناهمگنی تومور^۴ نامیده می‌شود. ناهمگنی تومور احتمالاً برای درمان سرطان و کشف نشانگر زیستی، به ویژه در روش‌های درمانی هدفمند، تأثیراتی خواهد داشت [۳۰]. درمان‌های فعلی، سرطان را به عنوان یک بیماری همگن درمان می‌کنند [۷۴].

داروهای هدفمند در برابر زیرجمعیت‌های تک یا چند سلولی با انکوژن^۵ جهش یافته که آن‌ها را هدف قرار می‌دهند، تولید شده اند، در حالی که آن دسته از زیرجمعیت‌های سلولی که هیچ گونه تاثیری از داروهای به واسطه جهش خود، نمی‌گیرند بدون درمان باقی مانده و ممکن است منجر به عود مجدد تومور یا عدم درمان تومور می‌شوند [۳۰]. این زیرجمعیت‌های سلولی بدون درمان ممکن است منجر به پیشرفت تومور پس از درمان دارویی شوند [۳۰]. به عنوان مثال، رشد مجدد سلول‌های تومورزا در سرطان روده بزرگ^۶ سرطان پستان و گلیوبالستوم^۷ پس از تابش یا درمان سیکلوفسفامید مشاهده شده است [۷۴]. بنابراین، مطالعه روند رشد تومور و ناهمگنی آن

¹Tumor

²Malignant tumor

³Cancer

⁴Tumor heterogeneity

⁵Oncogene

⁶Colorectal carcinoma

⁷Glioblastomas

تأثیرات زیادی بر تشخیص و درمان سرطان دارد.

تومورها می‌توانند خوش خیم، بد خیم و دارای رفتاری نامشخص یا ناشناخته باشند [۲]. تومورهای خوش خیم شامل فیبروییدهای رحمی^۸ و خالهای ملانوسیتیک^۹ است. آن‌ها محدود و محلی^{۱۰} هستند و به سرطان تبدیل نمی‌شوند [۴]. تومورهای بالقوه بد خیم^{۱۱} شامل سرطان در محل^{۱۲} هستند. آن‌ها به سایر بافت‌ها حمله نکرده و از بین نمی‌روند اما ممکن است به سرطان تبدیل شوند [۳]. تومورهای بد خیم را معمولاً سرطان می‌نامند. آن‌ها به بافت اطراف حمله کرده و از بین می‌روند، ممکن است متاستاز^{۱۳} ایجاد کنند و اگر درمان نشوند یا به درمان پاسخ ندهند، کشنده خواهد بود [۳].

ناهمگنی تومور توضیح می‌دهد که تومور بیش از یک نوع سلول شامل می‌شود. انواع مختلف سلول‌های داخل تومور دارای ویژگی‌های مورفولوژیکی و فیزیولوژیکی متمایزی مانند گیرنده‌های سطح سلول، تکثیر^{۱۴} و رگ‌زایی^{۱۵} هستند. ناهمگنی تومور می‌تواند بین تومورها (ناهمگنی بین توموری) و یا درون تومورها (ناهمگنی درون توموری) رخ دهد. به طور گستردگی پذیرفته شده است که توسعه تومور یک روند تکاملی است [۱۲]، و پیشرونده^{۱۶} معمولاً از یک سلول منشأ می‌گیرند و گروهی از سلول‌ها را تشکیل می‌شوند که در نهایت یک توده را شکل می‌دهند.

دو مدل برای ناهمگنی تومور وجود دارد (شکل ۱.۱). یک مدل تشکیل سرطان از طریق سلول‌های بنیادی بوده که قابلیت ارثبری ندارند و مدل دیگر تشکیل سرطان از طریق تکامل کلونی^{۱۷} بوده که قابلیت ارثبری دارد. [۱۲]. مفهوم سلول‌های بنیادی سرطانی بیان می‌کند که رشد و پیشرفت بسیاری از تومورها توسط کسری کمی از سلول‌ها کنترل می‌شود و اکثر سلول‌های موجود در تومور محصولات تمایز غیر طبیعی سلول‌های بنیادی سرطانی هستند [۱۲]. بنابراین، برای توصیف و از بین بردن سلول‌های بد خیم در تومورها، لازم است که بر بخش کوچکی از سلول‌های تومورزا تمرکز کنیم [۳۹]. مفهوم تکامل کلونی بیان می‌کند که تومور از یک سلول طبیعی ژنتیکی بوجود می‌آید که به تعداد زیادی سلول تبدیل می‌شود. در این تکامل، جهش‌های تصادفی به طور مداوم تولید می‌شوند و

⁸Uterine fibroid

⁹Melanocytic nevi

¹⁰Local

¹¹Potentially malignant tumor

¹²Carcinoma In Situ

¹³Metastases

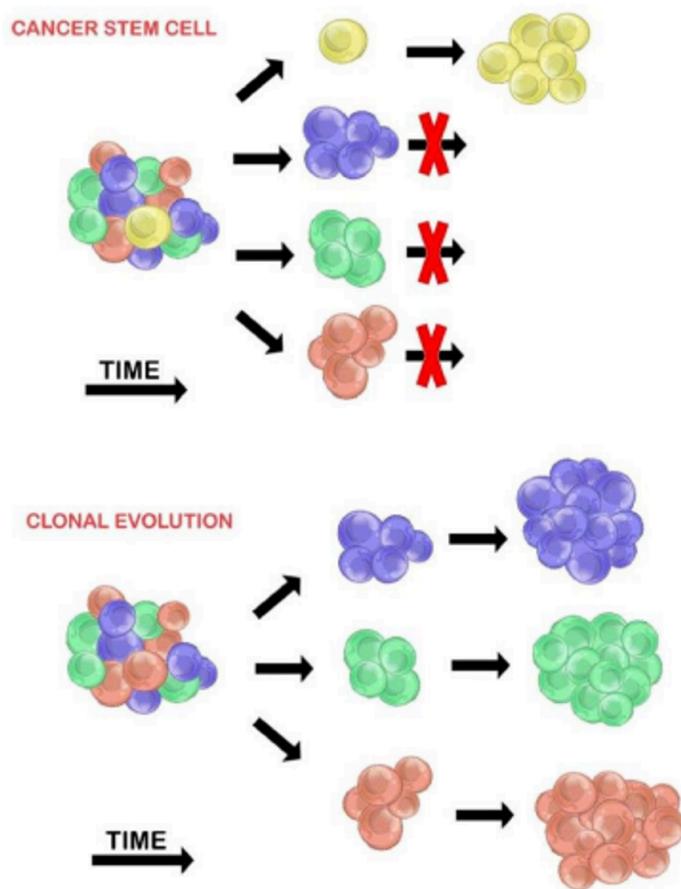
¹⁴Proliferative

¹⁵Angiogenic

¹⁶Spontaneous

¹⁷Clonal

در نهایت تومور حاصل میلیاردها سلول بدخیم است که حاصل از تجمع تعداد زیادی جهش است [۲۶]. تکامل تومور به عنوان توالی پیدرپی گسترش کلونی توصیف می‌شود، که در آن در هر حالت جدید یک رویداد جهش اضافی ایجاد می‌شود [۱۲].



شکل ۱.۱: دو مدل برای ناهمگونی تومور

یکی از توالی‌های پی در پی گسترش کلونی، یک مدل خطی از جانشینی کلونی است، جایی که جهش‌های متوالی پیدرپی باعث ایجاد توالی خطی از مجموعه‌های گسترش کلون می‌شوند و منجر به رشد کلون می‌شوند [۱۲]. مورد دیگر یک مدل چند کلونی از پیشرفت تومور است، که در آن یک سلول منفرد از طریق مکانیزم تقسیم به چندین زیرکلون گسترش می‌یابد [۴۷]. این مدل بیش از مدل خطی با ناهمگونی تومور مرتبط است. جهش‌های اکتسابی منجر به افزایش بی ثباتی ژنومی با هر نسل متوالی می‌شود [۱۷].

تومرهای ناهمگن^{۱۸} که مشکل از چندین کلون هستند، می‌توانند حساسیت‌های مختلفی را نسبت به داروهای سمیت سلولی^{۱۹} در نشان دهند. علاوه بر این، می‌زان ناهمگنی تومور می‌تواند خود به عنوان نشانگر زیستی^{۲۰} مورد استفاده قرار گیرد زیرا هر چقدر می‌زان ناهمگنی تومور بیشتر باشد، احتمال حضور کلون‌های مقاوم در برابر درمان بیشتر است [۷۷]. دلایل حساسیت‌های مختلف می‌توانند تعاملات بین کلون‌ها باشد که ممکن است اثر درمانی را مهار یا تغییر دهد [۱۲]. تومورهایی با ناهمگنی زیاد، با احتمال بیشتری از کلون‌های گوناگون تشکیل شده است که به درمان مقاوم هستند و ممکن است منجر به عدم موفقیت در درمان شوند. روش‌های نوین درمان تومورها با هدف شخصی‌سازی برنامه‌های درمانی از طریق هدف قرار دادن جمعیت‌های سلولی توموری موجود در یک بیمار، توسعه می‌یابند [۲۹]. ناهمگنی‌های توموری بکی از عوامل اصلی مقاومت در برابر دارو است و بنابراین، یک عامل بالقوه در شکست درمان محسوب می‌شود. [۲۹]. تومورها می‌توانند از راههای مختلف به طور همزمان به مقاومت دارویی دست یابند، بنابراین هدف قرار دادن فقط یک مکانیسم مقاومت برای غلبه بر نارسایی درمانی، می‌تواند مزیت درمان‌های هدفمند را محدود کند [۱۴]. بنابراین، ناهمگنی تومور می‌تواند برای درک توسعه تومور، پیچیدگی ایجاد کند و توسعه روش‌های موفقیت آمیز را با چالش رو برو کند [۲۹]. مطالعه ناهمگنی تومور می‌تواند منجر به پیشرفت و توسعه روش‌های درمانی شخصی‌سازی شده شوند و درک ما را از روابط عملکردی بین کلون‌ها در طول درمان افزایش دهند [۱۴]. برای مطالعه ناهمگنی تومور، بسیاری از ابزارهای محاسباتی موثر برای تجزیه و تحلیل اطلاعات کلونی تومور و تاریخچه تکامل آن تولید شده است. این ابزارها با استفاده از داده‌های تغییرپذیری ژنتیکی، تولید شده توسط فناوری‌های توالی یابی نسبتاً دقیق، قادر هستند تا ترکیب‌های کلونی تومور و رابطه اجداد بین کلون‌ها نتیجه دهند. این اطلاعات برای درک پیشرفت تومور و کمک به پیشرفت‌های درمانی کارآمد مهم است.

در ادامه مفاهیم حوزه تحقیق مثل مدل‌های ناهمگنی توموری، روش‌های مختلف توالی یابی، روش‌های مختلف ساخت درخت فیلورژنی تومور، مباحث مرتبط به یادگیری عمیق و یادگیری تقویتی به اختصار توضیح داده شد. در فصل سوم تحقیق پیشرو، به بررسی الگوریتم‌هایی که با استفاده از داده‌های توالی یابی تکسولی، درخت فیلورژنی تومور را استنباط کرده‌اند پرداخته شد. هر یک از این روش‌ها برای ساخت درخت فیلورژنی به همراه دادگان مورد استفاده، مورد ارزیابی قرار گرفت و در انتها فصل سوم مقایسه‌های بین روش‌های مختلف صورت گرفت. در فصل چهارم روش پیشنهادی استنباط درخت فیلورژنی بر مبنای یادگیری تقویتی و داده‌های

¹⁸Heterogenetic

¹⁹Cytotoxic

²⁰Biomarker

توالی یابی تکسولی به تفصیل بیان شده و در فصل پایانی نتایج بدست آمده و مقایسه آن با نتایج پیشین، گزارش شده است. در پایان موضوعات پیشنهادی که در کارهای آتی در راستای ادامه این پژوهش می‌تواند مورد بررسی قرار گیرند، توضیح داده شد.

فصل ۲

مبانی تحقیق

در این فصل ابتدا مفاهیم مورد نیاز جهت تعریف مسئله مانند مدل‌های ناهمگنی تومور، روش‌های یافتن درخت تکاملی تومور، روش‌های توالی‌یابی داده مورد بررسی قرار می‌گیرند. در ادامه مدل‌های مورد استفاده برای استنباط درخت تکاملی تومور معرفی می‌شوند. در پایان مفاهیم مرتبط با یادگیری ماشینی، یادگیری عمیق و یادگیری تقویتی به منظور استنباط درخت تکاملی تومور با رویکرد مبتنی بر داده^۱ توضیح داده می‌شوند.

۱.۲ تنوع ژنتیکی

دی‌ان‌ای^۲ یک مولکول بیولوژیکی است که توسط نوکلئوتیدها^۳ پلیمری شده است. در دی‌ان‌ای چهار نوع نوکلئوتید وجود دارد: آدنین^۴ (A)، تیمین^۵ (T)، سیتوزین^۶ (C) و گوانین^۷ (G). دی‌ان‌ای اساس توالی اسیدهای آمینه است که پروتئین را تشکیل می‌دهد. یک مولکول دی‌ان‌ای از دورشته تشکیل شده است. که در موازات^۸ هم و درجهت‌های مخالف قرار دارد و ساختاری از مارپیچ دوتایی ایجاد می‌کنند. هر نوع نوکلئوتید روی یک رشته

¹Data driven

²DNA

³Nucleotid

⁴Adenine

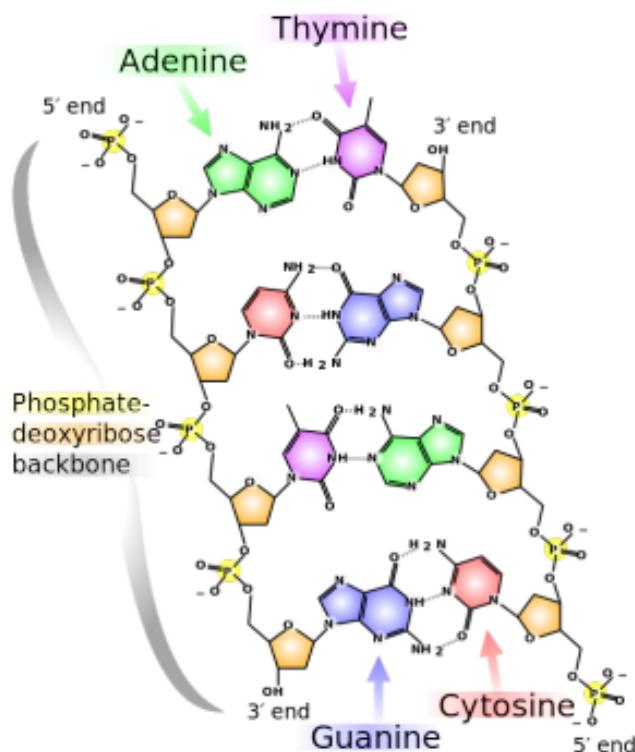
⁵Thymine

⁶Cytosine

⁷Guanine

⁸Antiparallel

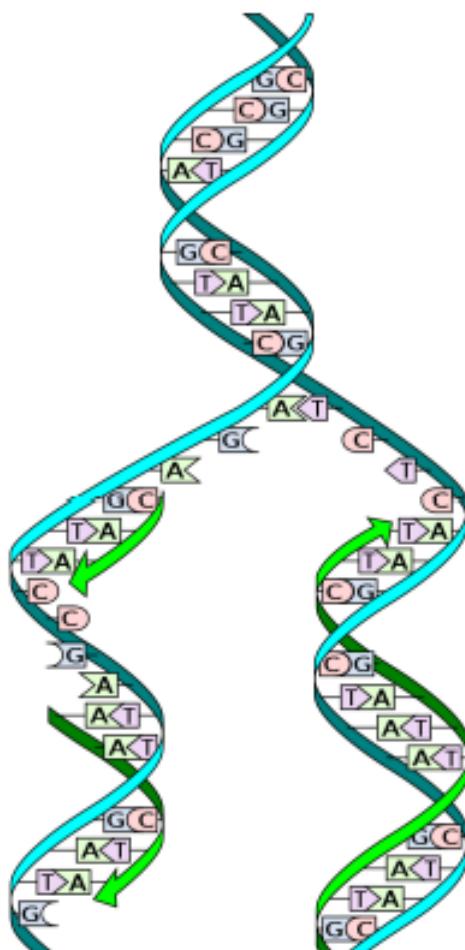
با نوع دیگری از نوکلئوتید در رشته دیگر مرتبط است: A با T؛ C با G (شکل ۱.۲) [۶]. این به عنوان قانون پایه جفت شدن نوکلئوتیدها در هر رشته از دی ان ای شناخته می شود.



شکل ۱.۲: مارپیچ دوگانه دی ان ای

همانند سازی دی ان ای فرآیند تولید دو مولکول دی ان ای یکسان از مولکول دی ان ای اصلی است. وقتی تکثیر شروع می شود، دو رشته یک مولکول دی ان ای از یکدیگر جدا می شوند و هر رشته به عنوان الگویی برای ساخت نمونه مشابه خود عمل می کند. نوکلئوتیدها در هر موقعیت از یک رشته با نوع دیگری از نوکلئوتید مبتنی بر قانون پایه جفت شدن، به منظور سنتز همتای این رشته، متصل می شود. پس از همانند سازی، مولکول دی ان ای اصلی به دو مولکول یکسان تبدیل می شود (شکل ۲.۲) [۶].

ژن ناحیه ای از دی ان ای است و به عنوان مولکول واحد وراثت شناخته می شود. ژن های متعددی در ساختار دی ان ای با عملکردهای متفاوت وجود دارد. جهش به تغییر دائمی توالی هسته ای ژنوم اتصال می شود. جهش ها می توانند در حین فرآیند تکثیر دی ان ای و با جفتگیری اشتباه در قسمت های مختلف دی ان ای ایجاد می شود.



شکل ۲.۲: همانندسازی دی ان ای

انواع مختلفی از جهش‌ها مانند جهش تک نوکلئوتیدی^۹ (جهش نقطه‌ای^{۱۰}) (شکل ۳.۲) و تغییرات ساختاری^{۱۱} شامل درج^{۱۲}، حذف^{۱۳} و برگشت^{۱۴} (شکل ۴.۲) وجود دارد. جهش‌های سلولی می‌توانند به بنا بر دلایلی چون مواد شیمیایی، سمیت یا ویروس ایجاد شوند. جهش در یک زن می‌تواند محصولات آن را تغییر دهد (مانند ایجاد پروتئین متفاوت) یا از عملکرد صحیح زن جلوگیری کند [۶].

⁹Single nucleotide mutation¹⁰Point mutation¹¹Single variant¹²Insertion¹³Deletion¹⁴reversion

original sequence:

ACTTGGTCA**G**AATTCCCAGGTGTCA

point mutation:

ACTTGGTC**A**TAAATTCCCAGGTGTCA

شکل ۳.۲: جهش تکنوکلئوتیدی

insertion:

ACTTGGTCA G AATTCCCAGGTGTCA
↓
ACTTGGTCAG**ATAGGC**AATTCCCAGGTGTCA

deletion:

ACTTGGTC**AGAATT**CCCAGGTGTCA
ACTTGGTCACCCAGGTGTCA

reversion:

ACTTGGTC**AGAATT**CCCAGGTGTCA
ACTTGGTC**TTAAGA**CCCAGGTGTCA

شکل ۴.۲: تغییرات ساختاری

۲.۲ تکامل تومور^{۱۵}

جهشی که در هر سلول از بدن اتفاق می‌افتد، به استثنای سلول‌های جنسی (اسپرم و تخمک)، جهش جسمی^{۱۶} نامیده می‌شود [۱]. تجمع جهش بدنی در طول زندگی یک فرد می‌تواند منجر به رشد کنترل نشده مجموعه‌ای از سلول (تومور) شود [۵۶] و می‌تواند باعث شکل‌گیری سرطان یا بیماری‌های دیگر شود [۱]. بدلیل تجمع سلول‌های گوناگون، بیش از یک نوع سلول در تومور وجود خواهد داشت. به گروههای سلول با مجموعه‌ای از

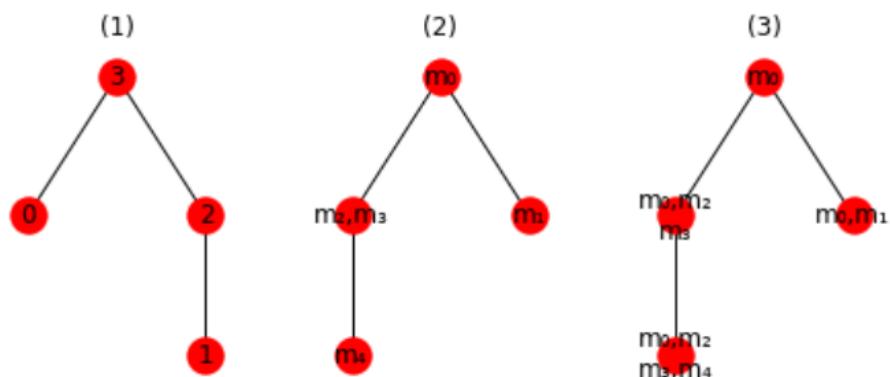
¹⁵Tumor Evolution

¹⁶somatic

جهش مشخص، کلون یا جمعیت سلولی تومور گفته می‌شود. کلون‌های موجود در تومور از نظر فیلوزنیک با هم مرتبط هستند و رابطه آنها را می‌توان با یک درخت فیلوزنیک نشان داد [۱۲]. درخت فیلوزنیک رابطه تکاملی بین کلون و ترتیب وقوع هر جهش را نشان می‌دهد. به عنوان مثال، شکل ۵.۲:

- یک درخت فیلوزنیک از یک تومور با چهار کلون با برحسب ۰ تا ۳ را نشان می‌دهد.
- جهش جدیدی را نشان می‌دهد که در هر کلون در طول تکامل این تومور رخ داده است.

همچنین هر کلون جهشی را در مسیر از کلون بالایی به سمت خود به ارث می‌برد. به عنوان مثال، کلون ۰ جهش‌های m_1 و m_0 دارد. کلون ۱ دارای جهش m_2 ، m_3 و m_4 است.



شکل ۵.۲: درخت فیلوزنیک تومور

۳.۲ تکنولوژی‌های توالی‌یابی و فراوانی تغییرات آلل^{۱۷}

تعیین توالی دی‌ان‌ای روشنی برای تشخیص ترتیب دقیق نوکلئوتیدها در یک رشته دی‌ان‌ای است. روش توالی‌یابی نسل بعدی^{۱۸} از تعدادی فناوری مدرن توالی تشکیل شده است که امکان تعیین هزینه و زمان توالی‌یابی را به طور موثر فراهم می‌کند. با استفاده از نمونه بیولوژیکی به عنوان ورودی این تکنولوژی‌ها، توالی‌های کوتاه نوکلئوتیدی تولید می‌شود (که به آن خوانش^{۱۹} گفته می‌شود). سپس خوانش با استفاده از الگوریتم هم‌ترازی^{۲۰}

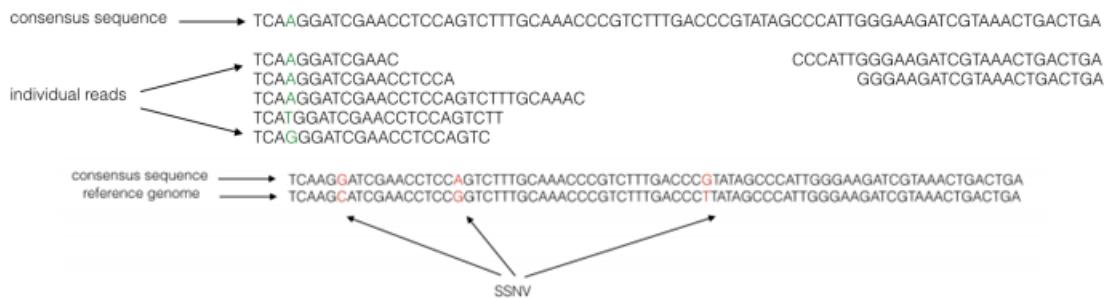
¹⁷Variant allele frequency

¹⁸Next generation sequencing

¹⁹Read

²⁰Alignment

متنوعی مانند الگوریتم تبدیل Burrows-Wheeler با ژنوم مرجع تراز می‌شوند. پس از ترازبندی، می‌توان با جمع‌آوری خوانش‌های همپوشانی^{۲۱}، توالی اجماعی^{۲۲} ایجاد کرد (شکل ۶.۲). در موقعیتی از توالی اجماع به دلیل همپوشانی خوانش‌ها، ممکن است بیش از یک نوع خوانش از نوکلئوتید تراز شده وجود داشته باشد (تعداد کل قرائت مرتبط با یک نوع جهش، را پوشش خوانش^{۲۳} نامیده می‌شود). نوکلئوتید موجود در این موقعیت به عنوان رایج‌ترین نوکلئوتید تراز شده، مشخص می‌شود. به عنوان مثال، در شکل ۶.۲، سه آدنین، (A) یک گوانین (G) و یک تیمین (T) در موقعیت سوم توالی اجماع تراز می‌شوند، سپس نوکلئوتید در آن موقعیت به عنوان آدنین (A) تعیین می‌شود. پس از ایجاد توالی اجماع، نوکلئوتیدهای موجود در آن توالی، که متفاوت از ژنوم مرجع هستند، شناسایی شده و به عنوان تغییرات بدنی تک نوکلئوتیدی^{۲۴} شناخته می‌شود. با استفاده از نمونه‌های متعدد استخراج شده از یک نمونه تومور، ما می‌توانیم تغییرات بدنی تک نوکلئوتیدی را در هر نمونه با فناوری تعیین توالی‌بایی تشخیص دهیم. نسبت تعداد سلول‌های موجود در یک نمونه حاوی تغییرات بدنی تک نوکلئوتیدی به کل سلول‌ها، فراوانی تغییرات آلل یک تغییر بدنی تک نوکلئوتیدی در این نمونه نامیده می‌شود. مقادیر فراوانی تغییرات آلل برای هر تغییر بدنی تک نوکلئوتیدی در هر نمونه تومور قابل محاسبه است. ابزارهای زیادی برای بازسازی درخت فیلورژنیک تومور از مقادیر فراوانی تغییرات آلل تومور به عنوان ورودی الگوریتم استفاده می‌کنند.



شکل ۶.۲: تشخیص تغییر بدنی تک نوکلئوتیدی از طریق خوانش هم‌ترازی

²¹Overlapping read

²²Consensus

²³Read coverage

²⁴Somatic single nucleotide variation

۴.۲ ناهمگنی ژنومی تومور

سرطان بیماری‌ای است که بدلیل ایجاد ناهنجاری‌های اساسی در فرآیندهای بنیادی سلول مانند تکثیر^{۲۵}، تمایز^{۲۶} و مرگ^{۲۷} سلول ایجاد می‌شود [۳۸]. این ناهنجاری منجر به رشد کنترل نشده تومور و به‌کارگیری بافت غیرسرطانی برای حمایت از این رشد می‌شود. علت اصلی این تغییرات جهش است. جهش یک اصطلاح گسترده است که چندین دسته از تغییرات ژنتیکی را پوشش می‌دهد. هنگام حاملگی، یک جنین دارای یک ژنوم خاص و منحصر به فرد است. این ژنوم که به ژنوم جوانه‌زنی^{۲۸} معروف است، می‌تواند با ژنوم انسانی مرجع مقایسه شود. ژنوم انسانی مرجع یک نمونه از ژنوم انسان است و از دی‌ان‌ای چند نفر تشکیل شده است. تفاوت بین ژنوم جوانه‌زنی و ژنوم مرجع به عنوان جهش ژنوم جوانه‌زنی شناخته می‌شود. جهش‌های جوانه‌زنی می‌توانند مسئول افزایش خطر ابتلا به سرطان باشند [۷۱]، اما بندرت خود مسئول مستقیم توسعه تومور هستند.

معمولًاً تومورها در اثر جهش‌های اکتساب شده پس از لقاح، که معروف به جهش‌های بدنی هستند، ایجاد می‌شوند. جهش‌های بدنی نتیجه اشتباهات در تکثیر دی‌ان‌ای [۱۱]، قرار گرفتن در معرض جهش‌های با منشأ داخلی یا خارجی یا واردشدن توالی‌های دی‌ان‌ای با منشأ بیرونی بدلیل قرار گرفتن در معرض ویروس است [۷۶]. غالباً در سرطان، جهش‌های بدنی باعث ایجاد اختلال در روند تکثیر دی‌ان‌ای یا ترمیم آن می‌شوند و حتی جهش‌های بدنی بیشتری ایجاد می‌کنند [۷۲]. نظریه کلونی بودن سرطان [۵۶] سرطان را به عنوان یک تک سلولی با منشأ غیرجنسي در نظر می‌گیرد که در اثر تولید مثل فراوان، یک توده متشكل از کلون‌های سلولی گوناگون را ایجاد می‌کند. در این مدل سلولهای توموری با یکدیگر در رقابت هستند و جهش‌های بدنی که مزیت رشد را ایجاد می‌کنند در جمعیت سلول‌های توموری از نسبت بیشتری برخوردار خواهند بود. جهش‌های بدنی که باعث رشد تومور شده و از سلولی به سلولی دیگر منتقل می‌شوند به عنوان جهش‌های راننده^{۲۹} شناخته می‌شوند. اولین سلولی که دارای جهش راننده بوده و آن را به جهش‌های بعدی منتقل می‌کند به عنوان سلول بنیانگذار شناخته می‌شود. همه فرزندان این سلول بنیانگذار، جهش راننده و هر جهش دیگری را که سلول بنیانگذار قبل از به دست آوردن جهش راننده بدست آورده است، دارند. این جهش‌های دیگر، که مزیتی برای رشد و گسترش تنوع توموری

²⁵Replication

²⁶Differentiation

²⁷Death

²⁸Germline genome

²⁹Driver mutation

ندارند، به عنوان جهش‌های مسافر^{۳۰} شناخته می‌شوند. شایان ذکر است که تعریف جهش رانده و مسافر به زمینه ژنتیکی و محیطی بستگی دارد. به عنوان مثال، شیمی درمانی داروهای سمیت سلولی (سیتوتوکسیک) می‌تواند باعث تغییر جهش از مسافر به جهش رانده شود و عامل اصلی مقاومت در برابر درمان باشد. همچنین جهش‌ها را می‌توان بر اساس نوع تغییری که در دی‌ان‌ای ایجاد می‌شود، به طبقات متمازی تقسیم کرد. حذف و تغییر تکنولوژیدها^{۳۱} جهش‌هایی هستند که یک پایه در ژنوم را به پایه دیگری تغییر می‌دهند. ایندل^{۳۲} درج یا حذف یک بخش دی‌ان‌ای است که می‌تواند کوتاه یا طولانی باشد. از ایندل کوتاه و تغییرات تک نوکلئوتیدی در مجموع به عنوان جهش‌های ساده بدنه^{۳۳} یاد می‌شود. در همه قسمت‌های یک ژنوم، از جمله کل کروموزوم‌ها، قابلیت حذف یا کپی شدن قسمتی از ژنوم وجود دارد. تغییرات شماره کپی به جهشی اتلاع می‌شود که منجر به حذف یا کپی شدن قسمتی از ژنوم می‌شود. تغییرات شماره کپی^{۳۴} نوعی تغییر ساختاری هستند که شامل وارونگی (وقتی قسمت بزرگی از ژنوم معکوس شده باشد) و انتقال متعادل (جایی که دو بخش ژنومی مکان‌های خود را با یکدیگر تعویض می‌کنند) می‌باشند^[۷۲]. این گونه‌های مختلف جهش مستقل از یکدیگر نیستند و می‌توانند در رابطه با یکدیگر اتفاق بیفتد (به عنوان مثال یک جهش می‌تواند منجر به تقویت یک وارونگی شود).

تکنیک توالی‌یابی نسل بعدی این امکان را فراهم کرده است تا با صرف هزینه بسیار کم و با استفاده از یک نمونه توموری، توالی‌یابی از دی‌ان‌ای صورت پذیرد و همین امر منجر به تحول گسترهای در زمینه مطالعه تکامل تومور شده زیر امکان نمونه-برداری در تعداد بسیار بالا از تومور فراهم می‌کند. نمونه‌گیری در حجم بالا این امکان را فراهم آورده است تا ناهمگنی تومور از نقطه منظر ژنتیکی مورد بررسی قرار گیرد و پاسخ به درمان بیماران سرطانی با جزئیات بیشتری مورد ارزیابی قرار گیرد.

نحویاً^{۳۵} همه نمونه‌های استخراج شده از تومور ترکیبی از سلول‌ها با ژنتیک‌های مختلف را شامل می‌شود. یک نمونه توموری به ندرت فقط شامل بافت سرطانی است زیرا شامل سلول‌های غیر سرطانی از استرومای اطراف^{۳۶} یا سلول‌های ایمنی نفوذی^{۳۶} است. مطالعات ژنومیک نشان داده است که حتی در میان سلولهای سرطانی، غالباً زیرجمعیت‌های متعدد سرطانی نیز وجود دارد. به عنوان مثال، در یک مطالعه مهم در سال ۲۰۱۲، گرلینگر و همکارانش^[۳۴] توالی‌یابی ژنوم و تغییرات شماره کپی را از طریق نمونه‌های مکانی مجزا استخراج شده از سرطان

³⁰Passenger mutation

³¹Single nucleotide variants (SNV)

³²Indel

³³Single Somatic Mutation

³⁴Copy number alteration

³⁵Surrounding stroma

³⁶Infiltrating immune cell

کلیه اولیه و نقاط متاستاز ثانویه بدست آورده‌اند. با بررسی این نمونه‌های متعدد، مشخص شد که یک ناهمگنی ژنتیکی قابل توجهی در تومور وجود دارد. تعداد بسیار زیادی از جهش‌های شناسایی شده در همه سلول‌های توموری مشاهده نشدند و این بدان معناست که این جهش‌ها بیش از آن‌که یک ناحیه کلونی باشند، به صورت یک ناحیه زیر کلونی بوده‌اند. با استفاده از روش‌های پردازش غیراتوماتیک، تغییرات تک نوکلئوتیدی‌ها و تغییرات شماره کپی بر اساس نمونه‌هایی که از آن استخراج شده‌اند، به خوش‌های مجزا دسته‌بندی شده و یک درخت فیلوزنی به آن‌ها نسبت داده شد. بازسازی درخت فیلوزنیک تومور این امکان را فراهم آورد تا سیر تکاملی تومور با استفاده از شاخه‌های مختلف درخت فیلوزنی شامل جهش‌هایی با عملکرد یکسان از سه ژن متفاوت مورد بررسی قرار گیرد.

در همان سال، یک مطالعه مهم دیگر، "تاریخچه زندگی ۲۱ سرطان پستان" [۵۵]، حضور ITH را نیز نشان داد. در این مطالعه آنها توالی‌یابی کامل ژنوم را در عمق متوسط ۱۸۸X بر روی تومور پستان a PD4120a انجام دادند. این عمق اجازه می‌دهد تا جمعیت‌های شیوع تا ۵٪ کم باشد. آنها مشاهده کردند که تغییرات تک نوکلئوتیدی‌ها در تعداد کمی از خوش‌های مجزا مشاهده می‌شوند که با توجه به کسر نوع آلل (VAF) آنها مشاهده می‌شود، نسبت خواندن‌ها در یک مکان متفاوت شامل آلل نوع. علاوه بر این، آنها توانستند نشان دهند که برخی از این خوش‌های مجزا را نمی‌توان با جهش‌های موجود در تمام جمعیت‌های سرطانی توضیح داد، که این نشان دهنده حضور تغییرات تک نوکلئوتیدی‌های تحت کلونال است. در همان زمان، آنها دریافتند که بسیاری از جهش‌ها در تمام سلول‌های سرطانی موجود در نمونه وجود دارد، که نشان می‌دهد جد مشترک اخیر نسبتاً دیر در زمان تکامل رشد کرده است. مشاهده اینکه جهش‌های زیر کلونال به جای توزیع یکنواخت یا مطابق قانون قدرت در خوش‌های متمایز پیدا شده است، شواهدی را نشان می‌دهد که این جهش‌های زیرکلونالی بیش از آنکه ناشی از تکامل خشی یا مصنوعات فنی باشد، در زیرمجموعه‌های متمایز ناشی از فشارهای انتخابی یافت می‌شود. نویسنده‌گان همچنین با تأیید اینکه جهش‌های زیر کلونال محدود به تغییرات تک نوکلئوتیدی نیستند، توانستند حضور تغییرات شماره کپی‌های کلونال و زیرکلونال را تأیید کنند. نویسنده‌گان یک الگوریتم خوش‌بندی غیر پارامتریک (یک مدل مخلوط فرآیند دیریشله (DPM)) را با استدلال قابل توجه دستی برای استباط فیلوزنی شاخه‌ای از چهار زیر جمعیت سرطانی در آن نمونه منفرد تومور ترکیب کردند. درک معماری ژنتیکی این زیر جمعیت‌ها می‌تواند به مطالعه زیست‌شناسی سرطان کمک کند و نشان داده شده است که در پیش‌بینی بقا در بسیاری از انواع سرطان مفید است [۴]. به عنوان مثال، زیر جمعیت‌های مختلف، که توسط مجموعه جهش‌های جسمی حمل شده تعریف می‌شوند، توانایی‌های مختلفی در مقاومت در برابر درمان و متاستاز دارند. برای انجام این کار، باید از یک یا تعداد

کمی از نمونه‌های تومور فله، ژنوتیپ‌های موجود در نمونه را شناسایی کرد. این مسئله، تحت عنوان بازسازی ساب کلونال، موضوع اصلی این پایان‌نامه است. مطالعات پیشگام که نشان داد ITH برای انجام این بازسازی به استدلال دستی قابل توجهی نیاز دارد. استدلال دستی کند، مستعد خطا است و به تخصص قابل توجهی نیاز دارد. مزایای بازسازی کاملاً خودکار بدیهی است. این بخش پیش زمینه مشکل بازسازی زیر کلونال، چگونگی پرداختن به آن برای انواع مختلف جهش، خصوصیات اصلی الگوریتم‌های بازسازی زیر کلونال و خلاصه‌ای از کارهای موجود در این زمینه را توصیف می‌کند.

۵.۲ بازسازی زیر کلونال

بازسازی ساب کلونال سعی دارد ژنوتیپ‌های موجود در تومور را از تعداد کمی از نمونه‌های توالی دی‌ان‌ای از آن تومور استباط کند. تعداد ژنوتیپ‌های موجود در تومور از قبل مشخص نیست. این ژنوتیپ‌های زیر کلونال به طور معمول با جهش‌هایی که در مقایسه با ژنوم خط جوانه‌ای دارند، توصیف می‌شوند. ژنوم جوانه‌زنی علاوه بر نمونه‌(های) تومور، با تعیین توالی یک نمونه غیرسرطانی تعیین می‌شود. در حال حاضر در هنگام تعریف این جمعیت از دونوع جهش به طور معمول استفاده می‌شود: جهش‌های ساده بدنی‌های متتشکل از تغییرات‌ها و درج / حذف کوچک (ایندل) و تغییرات شماره کپی حاصل از تغییرات ساختاری بزرگتر. مشاهده انواع جهش‌های دیگر، مانند مجموعه گسترده‌ای از SV‌ها که شامل بازآرایی هستند، مشاهده آنها دشوارتر است و روش‌های شناسایی آنها در مراحل اولیه رشد است.

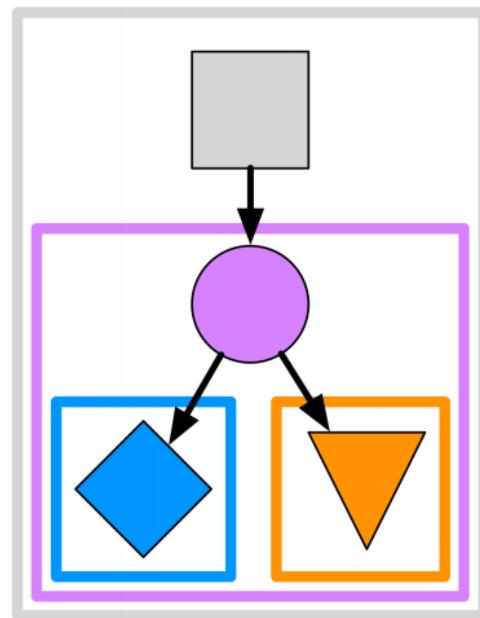
به طور متوسط، حتی در شرایط ایده‌آل، هر سلول در هر بخش یک جهش پیدا می‌کند [۱۱]، به همین ترتیب، بیشتر سلول‌های تومور ژنوتیپ منحصر به فردی خواهند داشت. بنابراین، به طور دقیق، اکثر سلول‌های تومور می‌توانند به طور بالقوه نمایانگر زیر جمعیت منحصر به فرد خود باشند. با این حال، به طور عملی، جهش‌هایی که مختص سلول‌های منفرد است یا فقط تعداد کمی از سلول‌ها آنها را به اشتراک می‌گذارد، در حین فراخوانی نوع شناسایی نمی‌شوند. تماس متغیر در بخش ۳.۵.۲ بیشتر مورد بحث قرار گرفته است. علاوه، سلول‌هایی که بخش عمده‌ای از جهش‌های خود را به اشتراک می‌گذارند، خصوصاً جهش‌های راننده، صفات مشابهی دارند. به همین ترتیب، من قرارداد گسترده‌ای را اتخاذ کرده و یک زیر جمعیت را به عنوان تمام سلول‌هایی که دارای زیر مجموعه یکسان جهش‌های بدنی در هنگام فراخوانی نوع هستند، تعریف می‌کنم.

یک گام مهم در بازسازی ساب کلونال محاسبه شیوع سلوی تبارهای زیر کلونال و سپس، در نهایت، زیر جمعیت‌های سرطانی است. شیوع سلوی یک زیر جمعیت، نسبت سلوهای نمونه توالي شده متعلق به آن است. غالباً، شیوع سلوی با تقسیم بر خلوص نمونه، یعنی نسبت سلوهای سرطانی در نمونه، به بخش سلوهای سرطانی، نسبت سلوهای سرطانی، تبدیل می‌شود. هر سلو دقيقاً به یک زیر مجموعه تعلق دارد، بنابراین این شیوع باید در یک جمع باشد. به طور کلی، سلوهای غیر سرطانی در یک زیر مجموعه واحد قرار می‌گیرند. با این حال، از آنجا که جهش‌ها اغلب در زیر جمعیت‌های متعدد وجود دارند، شیوع سلوی بسیاری از زیر جمعیت‌ها را نمی‌توان مستقیماً از جهش‌های آن استنباط کرد. برای پرداختن به این موضوع، ما یک نسب زیر کلونال برای یک جهش به عنوان مجموعه زیر جمعیت‌هایی که در آن وجود دارد، تعریف می‌کنیم. به طور رسمی، دودمانهای زیر کلونال از زیر جمعیت بنیانگذار تشکیل می‌شود (جایی که جهش برای اولین بار ظاهر می‌شود) و همه زیر جمعیت‌های بعدی آن (که وراثت جهش) علاوه بر جهش‌های خاص خود، این زیر مجموعه‌های فرزندی حاوی تمام جهش‌های موجود در نژاد تعریف کننده زیر جمعیت هستند (به جز در صورت حذف محل منبع جهش، برای جزئیات بیشتر به فصل ۳ مراجعه کنید). نسب مربوط به یک زیر درخت (یا کلاد) از درخت کلون تومور است. شیوع سلوی یک تبار مجموع شیوع سلوی زیر جمعیت‌هایی است که متعلق به آن تبار هستند. از آنجا که سلوهای می‌توانند در چندین نژاد زیر کلونال وجود داشته باشند، شیوع نسب در یک جمع نیست.

شکل ۷.۲ تصویری از یک درخت کلون نمونه را ارائه می‌دهد. گره‌های موجود در درخت، همانطور که در بالا تعریف شد، نشان دهنده زیر جمعیت است. فلش‌ها از جمعیت والدین به سمت فرزندانشان هدایت می‌شوند. دودمانهای زیر کلونال به صورت مستطیل نشان داده می‌شوند و با توجه به زیر مجموعه بنیادی آنها که در ریشه تیغه یافت می‌شوند، رنگی هستند.

۶.۲ تغییرات تعداد کپی

بیشتر ژنوم انسان دیپلوفید است، به این معنی که دو نسخه از توالي دی‌ان‌ای ما در سلوهای ما وجود دارد، یکی از پدر و دیگری از مادر. تغییرات شماره کپی این تغییر را می‌دهند، یا با تغییر در تعداد نسخه‌ها (مثلاً از طریق تکثیر کل ژنوم)، نسبت کپی‌های مادر به پدر (مثلاً از دست دادن خنثی هتروزیگوزیته در تعداد کپی‌ها، جایی که برای همان منطقه یک ژنوم والدین تکثیر می‌شود و دیگری حذف شده است) یا هر دو (به عنوان مثال



شکل ۷.۲: درخت کلون تومور

کپی کروموزوم مادر). بیشتر این تغییرات (به استثنای تکثیر کل ژنوم) دامنه محدودی از ژنوم را تحت تأثیر قرار می‌دهد، اما می‌تواند از تأثیر یک ژن تا یک کروموزوم کامل باشد. این بخش از ژنوم تغییر یافته به عنوان یک بخش شناخته می‌شود.

تغییرات شماره کپی می‌توانند تعداد کپی کل یک بخش و / یا تعداد نسبی نسبی دو کروموزوم والدین را تغییر دهند. هر یک از این تغییرات توسط توالی یابی ژنومی هسته قابل تشخیص است. تغییر در تعداد کپی کل یک بخش را می‌توان تشخیص داد زیرا نسبت خواندن آن نقشه به آن بخش بین خط جوانه زنی و نمونه تومور متفاوت خواهد بود. بخش از یک قطعه نسبت ورود خوانده شده است که به یک قطعه در یک نمونه غیر سرطانی ترسیم شده است به نسبت خوانده شده که به یک بخش در یک نمونه سرطانی ترسیم شده است. از نسبت نسبت‌ها برای محاسبه این واقعیت استفاده می‌شود که تعداد کل قرائت‌ها اغلب بین توالی یابی سرطانی و غیرسرطانی متفاوت است، در مناطق مختلف ژنوم عمق خواندن بیشتر یا پایین‌تر ناشی از محتوای GC یا نقشه برداری وجود دارد و تردستی یک تومور با بافت طبیعی متفاوت است. تکرر یک ژنوم، میانگین تعداد کپی از هر کروموزوم است که برای طول کروموزوم نرمال می‌شود.

با تغییر در کسر آلل می‌توان عدم تعادل در تعداد نسخه‌های مادری و پدری این بخش را تشخیص داد. در

مناطق دیپلوبید ژنوم‌ها، اگر یک بازه بین کپی‌های مادر و پدر متفاوت باشد، موقعیت هتروزیگوت نامیده می‌شود. جهش‌های تک پایه، خط جوانه زنی همچنین به عنوان چند شکلی تک هسته‌ای نامیده می‌شوند. وقتی یک ژنوم توالی‌یابی شود، حدود نیمی از قرائت آن مکان هتروزیگوت حاوی هر یک از بازها خواهد بود، در نتیجه کسر آلل ۵۰ است. این امر تا زمانی که نسبتی برابر با نسخه‌های مادرانه و پدری وجود داشته باشد، صادق خواهد بود. اگر این نسبت تغییر کند، کسر آلل تمام پولیمورفیسم تک هسته‌ای در بخش آسیب دیده تغییر می‌کند. پولیمورفیسم تک هسته‌ای هتروزیگوت به طور متوسط هر ۱۵۰۰ باز [۱۵] رخ می‌دهد و بنابراین برای بخش‌های طولانی بسیاری از پولیمورفیسم تک هسته ایی هتروزیگوت تحت تأثیر قرار می‌گیرند. توزیع کسر آلل S تمام پولیمورفیسم تک هسته‌ای در بخش، حالت دوگانه‌ای پیدا می‌کند که هر حالت نشان دهنده نسبت نسخه‌های آن بخش از هر والد است.

فراخوانی CNA چالش برانگیز است زیرا با مشاهده مستقل هر بخش، مسئله هنوز مشخص نشده است. حتی با فرض اینکه هر بخش فقط توسط یک CNA تحت تأثیر قرار گیرد، CNA موسوم به سه پارامتر (نسبت سلولهای حاوی CNA، تعداد کپی‌های مادر و تعداد کپی‌های پدری) وجود دارد و فقط دو مشاهده برای توضیح وجود دارد (و کسر آلل)

همه روش‌ها با فرض اینکه تعداد کمی از نژادهای زیرکلونال مسئول بیشتر یا تمام تغییرات شماره کپی هستند، این ابهام را برطرف می‌کنند. روشی که توسط الگوریتم باتبرگ [۵۵] به کار رفته است، به بیشتر تغییرات شماره کپی وابسته به یک نژاد زیر کلونال منفرد و شایع به نام تبار کلونال متکی است. تحت این روش، شیوع این تبار، همراه با تعداد کپی اصلی و جزئی در تمام تغییرات تعداد کپیکلونال، می‌تواند با یک فرآیند دو مرحله‌ای تخمین زده شود. در گام اول، این روش با فرض شیوع نژاد کلون f_c آغاز می‌شود. شیوع تبار کلونال در بیشتر موارد با خلوص نمونه تومور برابر است. با توجه به شیوع کلونال، هر بخش پس از آن فقط دو متغیر برای توضیح دارد (تعداد کپی بزرگ و جزئی). از آنجا که هر بخش دارای دو مشاهدات است، اکنون مسئله هنوز به درستی تعیین نشده است و بهترین کپی اصلی و مینور متناسب است. سپس، ترکیب کلی مقدار Φ فرض شده با ترکیب مناسب در تمام بخشها تعیین می‌شود. الگوریتم با بهینه سازی این تناسب بهترین مقدار Φ را انتخاب می‌کند. سپس برای هر بخش، شماره کپی اصلی و جزئی با بهینه سازی متناسب بودن قطعه با بهترین مقدار Φ انتخاب می‌شود. این روش فرض می‌کند که تمام تغییرات شماره کپی به نژاد کلونال تعلق دارند، که همیشه درست نیست. در مرحله بعدی، بخش‌هایی که حاوی تغییرات تعداد کپی‌ها هستند با جستجوی بخش‌هایی با اطلاعات مناسب ضعیف با استفاده از Φ استباط شده مشخص می‌شوند. در این بخش‌ها، روش به طور همزمان و مستقل از هر

بخش دیگر، عدد Φ و عدد کپی بزرگ و جزئی را استنباط می‌کند.

از آنجا که سه متغیر وجود دارد و تنها دو مشاهده وجود دارد، راه حل‌های بسیاری با تناسب داده برابر وجود دارد که از نظر زیست شناختی برای این تغییرات تعداد کپی زیر کلونال قابل قبول است. این ابهام با انتخاب راه حلی که نزدیکترین شماره به شماره نسخه طبیعی است برطرف می‌شود، اما تعدادی از موارد متدائل وجود دارد که این ابتکار عمل ناموفق است. سپس این روش‌ها انتساب تغییرات تعداد کپی زیرکلونال به دودمان و تمام استنباط‌های فیلوزنیک را برای روش‌های پایین دست رها می‌کنند.

رویکرد عمده دیگر این است که فرض کنیم همه تغییرات شماره کپی از تعداد کمی تبار ساب کلونال به وجود می‌آیند. الگوریتم‌هایی که از این روش استفاده می‌کنند به طور مشترک شیوع این نژادها و تعداد کپی بزرگ و جزئی را برای هر بخش استنتاج می‌کنند (به عنوان مثال THetA [۷۸، ۸۶] و TITAN [۸۶]). تعداد دودمانهای زیر کلونال معمولاً با استفاده از احتمال جریمه شده‌ای مانند معیار اطلاعات بیزی (BIC) یا انواع BIC تعیین می‌شود (به عنوان مثال BIC از THetA اصلاح شده با پارامتر مقیاس گذاری استفاده می‌کند [۸۶]). بنابراین این روش‌ها هم تغییرات شماره کپیرا فراخوانی می‌کنند و هم آنها را به دودمانهای زیرکلونال اختصاص می‌دهند. هیچ روش موجود این دودمان‌ها را در یک درخت فیلوزنیک قرار نمی‌دهد

۷.۲ جهش‌های ساده بدنی

جهش‌های ساده بدنی جهش‌های کوچکی هستند که می‌توانند مستقیماً از طریق توالی یابی و نسبت کروموزوم‌های موجود در نمونه حاوی آنها از تعداد قرائت‌های حاوی جهش و تعداد کل خوانده‌ها در آن مکان، مشاهده شوند. نسبت قرائت حاوی جهش به کل قرائت به عنوان VAF جهش شناخته می‌شود. جهش‌های ساده بدنی‌ها معمولاً با بررسی مشترک ترازاها و یک نمونه غیرسرطانی خوانده می‌شوند. این استنباط مشترک برای جداسازی انواع بدنی و ژرمنیال مورد نیاز است.

این فرایند به دلیل انواع مختلف خطاهای و تعصبات که در داده‌های NGS وجود دارد، دشوار می‌شود [۳۱]. یک مشکل اساسی در تشخیص جهش‌های ساده بدنی این است که به نظر می‌رسد خطاهای توالی جهش‌های ساده بدنی شیوع کمی دارند. به طور خاص، در Illumina Hiseq2000 که به طور گسترده استفاده می‌شود، از هر ۱۰۰۰ پایه یکی از آنها دارای یک خط است (به طور معمول یک تعویض) [۵۹]. به همین ترتیب، در طول سه

میلیارد پایه ژنوم انسانی، یک احتمال غیر قابل اغماض وجود دارد که در بعضی موقعیت‌ها، چندین بار خواندن دقیقاً شامل خطای توالی دقیقاً در همان موقعیت‌ها است. به نظر می‌رسد این خطاهای شیوع کم جهش‌های ساده بدنی دارند. تمایز بین این خطاهای شیوع کم واقعی جهش‌های ساده بدنی‌ها شامل یک معامله بین حساسیت و ویژگی و در حالت ایده‌آل، یک مدل نویز بسیار دقیق است. حل این مشکل امتداد طبیعی کار گستردگی است که در زمینه فراخوانی جهش‌های جوانه‌زنی انجام شده است و الگوریتم‌های زیادی برای انجام این کار وجود دارد (به عنوان مثال [۲۰، ۲۱])

۸.۲ ترک آللی^{۳۷}

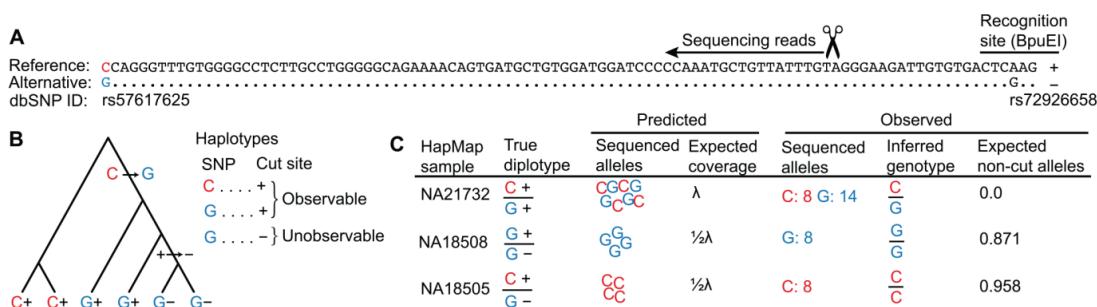
اگرچه روش‌های تعیین توالی با بازدهی بالا [۴۲] ارزان هستند، اما تحت تاثیر مقدار بایاس هستند و مارکرهای ژنتیکی‌ای تولید می‌کنند که تقریباً به طور تصادفی در کل ژنوم تقسیم می‌شوند. این روشها با موفقیت در نگاشت صفات [۳۵، ۵۶]، ساخت مپ پیوندی [۲۸، ۶۱]، اسکن انتخاب [۲۱، ۷۹]، و برآورد تنوع ژنتیکی [۱۹] استفاده شده است. یکی از این روش‌ها، تعیین ژنوتیپ براساس توالی [۷] (GBS) است. در GBS، هدف توالی‌یابی فقط با اتصال آداتورهای توالی به محل‌های برش آنزیم محدود کننده، به کمتر از ۵٪ از ژنوم کاهش می‌یابد (شکل زیر). قرائت GBS همچنین می‌تواند به صورت کانکت‌های کوتاه مونتاژ شود، که بدون نیاز به توالی ژنوم فراخوانی یک نوع تغییر تک هسته‌ای (تغییرات تک نوکلئوتیدی) را امکان پذیر می‌کند [۳۸]. از این رو، GBS یک روش محبوب در سیستم‌های غیر مدلی است که به طور معمول فاقد منابعی مانند مجموعه ژنوم و ریزآرایه‌ها است.

بر خلاف توالی‌یابی کل ژنوم (WGS)، GBS مستعد ابتلا به خطاهای مختلف تماس به دلیل محدودیت چندشکلی‌های سایت است (کاهش آللیک). کاهش آللیک در GBS می‌تواند برنامه‌هایی را که به فراخوانی دقیق تغییرات نادر، از جمله تخمین طیف فرکانس سایت در ژنتیک جمعیت متکی هستند، را دچار اختلال کند. یک رویکرد آماری سیستماتیک برای تشخیص کاهش آللیک در داده‌های توالی GBS، اجرا شده و در بسته نرم افزاری منبع باز GBStools وجود دارد. این روش مبتنی بر این واقعیت است که کاهش آللیک متناسب با تعداد آلل‌های سایت محدود کننده بدون برش که در آنجا حمل می‌کند، میزان خوانش نمونه را در یک سایت خاص کاهش می‌دهد. بنابراین GBStools پوشش هر نمونه را در یک سایت خاص به عنوان یک متغیر تصادفی پواسون مورد

³⁷ Allelic dropout

³⁸ Mapping

استفاده قرار می‌دهد که از توزیع با میانگین λ (آللیک‌های بدون برش صفر)، توزیع با میانگین $\frac{1}{2}\lambda$ (یک آللیک بدون برش)، یا با میانگین صفر (دو آللیک بدون برش). GBStools حداکثر احتمال پارامتر λ را با استفاده از تعداد واقعی آللیک‌های بدون برش در هر نمونه که به عنوان متغیرهای نهفته (مشاهده نشده) در نظر رفته می‌شود و از طریق حداکثر رساندن مقدار چشم انتظاری (EM)، محاسبه می‌کند. از مقادیر مورد انتظار این متغیرهای نهفته می‌توان برای تخمین اینکه کدام نمونه‌ها یک آللیک بدون برش دارند استفاده کرد. به طور همزمان، GBStools فرکانس سایت آلل‌های SNP مرجع قابل مشاهده و جایگزین، φ_1 و φ_2 ، و آللیک بدون برش، φ_3 ، که در آن $\varphi_3 = \varphi_1 + \varphi_2$ برآورد می‌کند و در نهایت، آزمون نسبت احتمال با مقایسه فرضیه صفر $= 0$ با فرضیه > 1 جایگزین می‌کند. GBStools در اجرای فعلی خود نمی‌تواند ژنتوتیپ‌های واقعی پنهان شده توسط کاهش آللیک را استباط کند، اما می‌توان با فیلتر کردن سایت‌هایی که نسبت احتمال آنها زیاد است خطاهای را حذف کند.



شکل ۸.۲: نمایی از تطابق ژنتیکی

در شکل بالا، آلل BpuEI بدون برش ناشی از SNP rs72926658 با برچسب “-” و آلل برش با “+” برچسب گذاری شده است. آلل “-” در هاپلوتیپ با آلل G مستقیم شده بوجود آمده و باعث شده تا برخی از آلل‌های G توسط GBS قابل مشاهده نباشند. نمونه‌های نشان داده شده دارای سه دیپلوتیپ هتروزیگوت است. نتایج توالی با پیش‌بینی‌ها مطابقت داشت و نمونه NA18505 به اشتباه هموزیگوت نامیده می‌شد، اما انتظار می‌رود تعداد آلل‌های کاهشی محاسبه شده توسط (0.958) GBStools با تعداد واقعی (۱) مطابقت داشته باشد، و آن را به عنوان یک تماس اشتباه احتمالی مشخص کند.

۹.۲ مقدمه‌ای بر مدل‌سازی احتمالی

وظیفه اصلی یادگیری ماشین، یادگیری از داده‌ها است، کاری که به عنوان استنباط شناخته می‌شود. برای یادگیری از داده‌ها، باید فرضیاتی را مطرح کرد. توصیف رسمی فرضیات صورت گرفته به عنوان یک مدل ذکر می‌شود. یک مدل احتمالی مفروضات ارائه شده را تعریف می‌کند که اطلاعات آموخته شده را با استفاده از متغیرهای تصادفی و توزیع‌های احتمال به داده‌های مشاهده شده پیوند می‌دهد. توزیع‌های احتمال توابع ریاضی هستند که یک رویداد را ورودی می‌کنند و احتمال آن واقعه را بیرون می‌آورند. توزیع احتمال می‌تواند تابعی بیش از واقعه باشد و این متغیرهای اضافی به عنوان پارامترهای توزیع شناخته می‌شوند^[۳۷]. رویکرد بیزی در یادگیری ماشین شامل استنباط احتمالی مقادیر پارامترهای منوط به مشاهدات است^[۳۸]. چهار مولفه دارد:

- احتمال: احتمال مشاهده داده‌ها است، مشروط به تنظیم پارامتر $P(\text{data} | \text{parameters})$
- پارامترهای احتمال
- پارامترهای قبلی
- داده‌های مشاهده شده

پارامترها خود مجموعه‌ای از متغیرهای تصادفی هستند که از توزیع قبلی $P(\text{parameters})$ گرفته شده‌اند، که باورهای ما را در مورد احتمال حالت‌های مختلف پارامتر در غیاب مشاهده مشاهده می‌کند. این اصطلاحات با استفاده از قانون بیز با هم ترکیب می‌شوند:

$$P(\text{parameters} | \text{data}) = P(\text{data} | \text{parameters}) * P(\text{parameters}) / P(\text{data}) \quad •$$

$$\text{Posterior} \propto \text{likelihood} * \text{prior} \quad •$$

پس زمینه توزیع پارامترهای مشروط به مشاهده داده‌ها است و خروجی اصلی استنتاج بیزی است. از توزیع پسین می‌توان برای انجام کارهایی مانند پیش‌بینی مشاهدات آینده استفاده کرد.

۱.۹.۲ زنجیره مارکوف مونت کارلو^{۳۹}

برای انجام استنتاج بیزی^{۴۰}، ما اغلب می‌خواهیم در توزیع پسین ادغام شده، پیش‌بینی کنیم یا خلاصه‌هایی پیدا کنیم، به عنوان مثال میانگین پارامتر پسین. به طور کلی، انجام چنین ادغامی (جمع بندی در مورد متغیرهای گستته) از نظر تحلیلی غیرقابل حل است. با این حال، می‌توان چنین ادغام‌هایی را با استفاده از نمونه‌هایی که از قسمت پسین ترسیم شده‌اند تقریبی داد:

$$E[f] = \int f(x)p(x)dx \approx 1/N \sum_{i=1..N} f(x_i) \quad (1.2)$$

که در آن x_i نمونه i از $p(x)$ و $f(x)$ به ترتیب توزیع و عملکرد مورد نظر ما است. به ندرت می‌توان مستقیماً از توزیع پسین نمونه برداری کرد. برای تولید موثر نمونه‌ها از توزیع، حتی در ابعاد بالا، می‌توان از تکنیک زنجیره مارکوف مونت کارلو استفاده کرد. زنجیره مارکوف مونت کارلو یک زنجیره مارکوف می‌سازد که در آن توزیع تعادل توزیع پسین است. سپس مقادیر زنجیره می‌تواند به عنوان نمونه از پسین با توجه به همگرایی کافی به توزیع تعادل مورد استفاده قرار گیرد. برای انجام زنجیره مارکوف مونت کارلو، تاز زمانی که بتوان $p(x)$ را محاسبه کرد، نیازی به محاسبه $p(x)$ نیست. این زنجیره مارکوف مونت کارلو را قادر می‌سازد تا از محاسبه ثابت‌های نرمال سازی، که اغلب غیرقابل حل هستند، خودداری کند. یک زنجیره مارکوف به عنوان یک سری متغیرهای تصادفی تعریف می‌شود که دارای ویژگی استقلال شرطی زیر هستند:

$$p(z^{N+1} | z^1..z^N) = p(z^{N+1} | z^N) \quad (2.2)$$

نمونه‌ای از الگوریتم زنجیره مارکوف مونت کارلو الگوریتم Metropolis-Hastings (MH) است [۴۰]. الگوریتم MH از حالت دلخواه Z^t شروع می‌شود. سپس یک حالت پیشنهادی z از توزیع پروپوزال $q(z|z^t)$ ترسیم می‌شود. این حالت پیشنهادی z با احتمال زیر پذیرفته می‌شود:

$$\min \left(1, \hat{p}(z^*) q(z^t | z^*) / \hat{p}(z^t) q(z^* | z^t) \right) \quad (3.2)$$

³⁹Markov Chain Monte Carlo (MCMC)

⁴⁰Bayesian

می‌توان نشان داد که الگوریتم MH تعادل دقیق را برآورده می‌کند و از این رو، $(x)p$ توزیع تعادل است [۱۳]. حالی که توازن دقیق برای اثبات اینکه در محدوده نمونه‌های بی‌نهایت زنجیره به توزیع مورد نظر همگراست کافی است، اما در عمل فقط تعداد محدودی از نمونه‌ها را می‌توان ترسیم کرد. واضح است که نمونه‌های ابتدای زنجیره، که از یک مکان دلخواه در فضای حالت شروع می‌شوند، بعید است از توزیع تعادل باشد. این نمونه‌ها به عنوان نمونه‌های سوختنی کنار گذاشته می‌شوند. هرچه همگرایی زنجیره مارکوف سریعتر باشد، نمونه‌های کمتری باید کنار گذاشته شوند و می‌توان از تعداد بیشتری برای محاسبه انتظارات استفاده کرد. با بررسی اثری از مقادیر مهم پارامتر یا احتمال همگرایی می‌توان نظارت کرد، اما این امر ممکن است چند حالت را از دست بدهد. متاسفانه دانستن اینکه آیا همگرایی حاصل شده است غیرممکن است، فقط گاهی اوقات می‌توان همگرایی را رد کرد [۳۳]. گذشته از همگرایی، یکی دیگر از خصوصیات اصلی یک زنجیره مارکوف میزان اختلاط زنجیره است. با توجه به n نمونه مستقل از توزیع، واریانس میانگین پارامتر برآورده σ_n است که σ انحراف استاندارد توزیع خلفی پارامتر است. نمونه‌های گرفته شده از زنجیره مارکوف مستقل نیستند، زیرا به وضعیت فعلی زنجیره بستگی دارند (یعنی فقط از نظر شرطی مستقل هستند). برای تخمین اندازه نمونه موثر یک زنجیره مارکوف، یعنی تعداد نمونه‌های مستقل با همان خطای استاندارد همان زنجیره، می‌توان از معادله زیر استفاده کرد:

$$ESS = \frac{n}{1 + 2 \sum_{j=0}^{\infty} \rho_j} \quad (4.2)$$

حاصل جمع بی‌نهایت محاسبه ESS را می‌توان با استفاده از برآورده پریودوگرام کوتاه تطبیقی Sokal [۶۹] تخمین زد.

۱۰.۲ یادگیری ماشین^{۴۱} و یادگیری تقویتی^{۴۲}

آنالیز داده‌های بالینی یک حوزه مهم تحقیقاتی در انفورماتیک، علوم کامپیوتر و پزشکی است که توسط محققان شاغل در دانشگاه‌ها، صنعت و مرکز بالینی انجام می‌شود. یکی از بزرگ‌ترین چالش‌ها در تجزیه و تحلیل داده‌های پزشکی، استخراج و تجزیه و تحلیل داده‌ها از تصاویر است. در چند سال اخیر روش‌های یادگیری

⁴¹Machine learning

⁴²Reinforcement learning

ماشین انقلابی بزرگ در بینایی کامپیوتر^{۴۳} به وجود آورده است که راه حل های جدید و کارآمدی را در مورد خیلی از مسائل و مشکلات موجود در آنالیز تصاویر که مدت زمان طولانی است حل نشده باقی مانده اند معرفی می کنند. برای اینکه این انقلاب وارد حوزه آنالیز تصاویر پزشکی شود شیوه و روش های اختصاصی ای باید طراحی شوند تا خاص بودن تصاویر پزشکی را در نظر گیرند. سیستم های کامپیوتراً هوشمند چندین دهه است که در دنیا جایگاه برجسته ای پیدا کرده اند. در حال حاضر، به خاطر تکنیک های جدید هوش مصنوعی^{۴۴}، قابلیت پردازش کامپیوتراً بالا و رشد گسترده تصویربرداری و ذخیره سازی دیجیتالی داده، کاربرد هوش مصنوعی در حال انتقال به حوزه های گوناگون می باشد. در حوزه پزشکی، سیستم های هوش مصنوعی به منظور آشکارسازی بیماری، پیش بینی و به عنوان استراتژی پشتیبان در تصمیم گیری بالینی در حال توسعه، کاوش و ارزیابی هستند. در زمینه سرطان سینه^{۴۵} از هوش مصنوعی به منظور آشکارسازی زودهنگام و تفسیر ماموگرامها^{۴۶} به منظور بهبود غربالگری سرطان پستان و کاهش تشخیص مثبت کاذب^{۴۷} استفاده می شود و این امکان فراهم شده است تا متخصصانی مانند رادیولوژیست ها^{۴۸} بتوانند بر اساس میلیون ها تصویر از بیماران قبلی که مشخصات مشابهی دارند، تصمیمات آگاهانه ای بگیرند. استفاده از هوش مصنوعی در شیوه های تشخیص سرطان سینه به مدالیته تصویربرداری^{۴۹} و همچنین تفسیر آسیب شناسی^{۵۰} نیز گسترش یافته است. یادگیری عمیق^{۵۱} که زیر شاخه ای از یادگیری ماشین می باشد یکی از تکنیک های هوش مصنوعی است که در انواع مختلفی از مسائل کلینیکی و پردازش تصاویر پزشکی شامل آشکارسازی^{۵۲}/شناسایی^{۵۳}، قطعه بندی^{۵۴} و تشخیص به کمک کامپیوتراً^{۵۵} به کار گرفته می شود. یادگیری عمیق مجموعه ای از الگوریتم های ماشین است که قادر به مدل سازی الگوها به طور مستقیم از داده های خام می باشد. الگوریتم های یادگیری عمیق از مجموعه ای از لایه های چندگانه با واحد های پردازنده غیرخطی برای استخراج و تبدیل ویژگی استفاده می کنند. هر لایه از خروجی لایه قبل به عنوان ورودی استفاده می کند. این مفهوم با بسیاری از روش های دیگر یادگیری ماشین که نیاز به استخراج ویژگی دارند متفاوت است. به

⁴³Computer Vision⁴⁴Artificial Intelligence (AI)⁴⁵Breast cancer⁴⁶Mammogram⁴⁷False positive⁴⁸Radiologist⁴⁹Imaging modality⁵⁰Pathology⁵¹Deep learning⁵²Detection⁵³Recognition⁵⁴Segmentation⁵⁵Computer-aided diagnosis

همین ترتیب این الگوریتم‌ها حتی در مسائلی که دانش بسیار کمی در موردشان وجود دارد، می‌توانند مورد استفاده قرار گیرند. اگرچه در دهه ۱۹۹۰ این الگوریتم‌ها در برخی از مطالعات مورد استفاده قرار گرفته‌اند، اما در چند سال اخیر شاهد نتایج بسیار چشمگیر این الگوریتم‌ها هستیم. با توجه به وجود داده‌های بیشتر و همچنین قدرت محاسباتی بالا، این روش‌ها در بسیاری از زمینه‌ها توانسته‌اند به عملکرد انسان یا بهتر از انسان دست یابند^[۵]. شبکه‌های عصبی مصنوعی نوع خاصی از مدل‌های یادگیری عمیق هستند که برای کار با داده‌های از نوع تصویر مناسب هستند.

شبکه‌های عصبی مصنوعی مدل‌هایی هستند که در بسیاری از زمینه‌های تحقیقاتی از جمله یادگیری ماشین کاربرد دارند. یک شبکه عصبی مصنوعی از واحدهای ساده‌ای به نام نورون^{۵۶} تشکیل شده است که در یک سیستم پیچیده سازمان یافته‌اند. هر نورون بر اساس ورودی‌های خود، یک خروجی (فعال‌سازی^{۵۷}) را محاسبه می‌کند که می‌تواند فعالیت‌ها یا داده‌های سایر نورون‌ها باشد. متدائل‌ترین نوع شبکه عصبی، شبکه عصبی کاملاً متصل شبکه عصبی کاملاً متصل پیش‌خور^{۵۸} است. این شبکه‌ها دارای ورودی (جایی که داده‌ها وارد می‌شوند) و خروجی هستند. به طور معمول، هدف از استفاده از این مدل‌ها حل رگرسیون^{۵۹} یا طبقه‌بندی^{۶۰}، توسط تقریب فعال‌سازی خروجی با مقدار هدف، برای هر داده ورودی است. این شبکه‌ها به صورت لایه^{۶۱} متوالی سازماندهی شده‌اند که یک نورون (واحد) از لایه k تمام نورون‌لایه $1 - k$ را به عنوان ورودی دریافت می‌کند، ترکیبی خطی از این مقادیر را محاسبه کرده و آن را از طریق تابع غیر خطی عبور می‌دهد

محاسبه خروجی نورون i ام لایه k

$$O_{k,i} = \text{actv}(W_{k,i} \cdot l_{k-1} + b_{k,i}) \quad (5.2)$$

که واحد i ام لایه k و l_{k-1} بردار تمام فعال‌سازهای لایه $1 - k$ است. بردار $W_{k,i}$ و عدد $b_{k,i}$ پارامترهای ما هستند که اغلب به آنها وزن شبکه^{۶۲} گفته می‌شود که برای یک وظیفه خاص آموخته می‌شوند. تابع فعال‌سازی غیرخطی actv می‌تواند اشکال مختلفی به خود بگیرد. هر مدل با یک لایه پنهان و تعداد مشخصی نورون اگر

⁵⁶Neuron

⁵⁷Activation

⁵⁸Fully-connected feed forward neural network

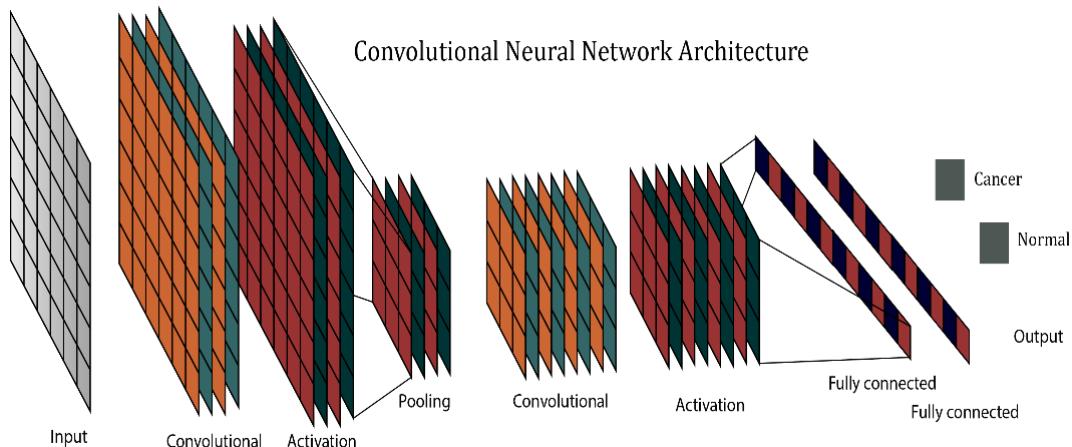
⁵⁹Regression

⁶⁰Classification

⁶¹Layer

⁶²Network weight

پارامترهای کافی داشته باشد می‌تواند هر تابع پیوسته‌ای را با خطاب دلخواه تقریب بزند [۲۳]. شبکه‌های عصبی کانولوشنی^{۶۳} یک نوع شبکه عصبی مصنوعی هستند که از نورون‌ها، لایه‌ها و وزن‌ها تشکیل شده‌اند. مطالعه‌ای که در سال ۱۹۶۸ میلادی صورت گرفت نشان داد که قشر بینایی مغز برای پردازش اطلاعات از تصاویر از الگوی پیچیده‌ای استفاده می‌نماید [۷۵]. نواحی ادراکی که قشر بینایی در آن قرار دارد، همانند فیلترهای محلی بر روی اطلاعات تصویر اعمال می‌شود. سلول‌های ساده‌تر برای تشخیص ویژگی‌های ادراکی سطح پایین‌تر در نواحی ادراکی مانند لبه‌ها کاربرد دارند، همچنین سلول‌های پیچیده قادر به تشخیص ویژگی‌های مهم‌تر و اختصاصی‌تر و در سطوح بالاتر می‌باشند. تشخیص ویژگی‌های اختصاصی‌تر نتیجه و ترکیبی از ویژگی‌های سطح پایین می‌باشد. این عملکرد مغز الهام بخش شبکه‌های عصبی عمیق امروزی می‌باشد. مفهوم شبکه کانولوشن نخستین بار در سال ۱۹۸۰ توسط فکوشیما مطرح گردید [۳۲]. اما به دلیل نیاز به سخت افزارها و پردازشگرهای گرافیکی قوی استفاده از این شبکه‌ها برای تشخیص تا سال ۲۰۱۲ که به شکل اختصاصی برای تشخیص تصاویر ارایه و معرفی گردید به تعویق افتاد [۴۶].



شکل ۹.۲: معماری یک شبکه عصبی کانولوشنی

همانطور که قبل^{۶۴} بیان شد، شبکه‌های عصبی کانولوشنی مدل‌های شبکه عصبی کاملاً متصل پیش‌خور هستند که از لایه‌های زیادی تشکیل شده‌اند. بسیاری از این مدل‌ها محدودیت‌های پارامتر و مکانی دارند که در ادامه توضیح داده خواهد شد. با این حال، آنها در تغییراتی که بر ورودی‌شان اعمال می‌کنند تفاوت دارند. در اینجا ما تمام لایه‌های یک شبکه کانولوشنی و توابع مورد استفاده در آموزش آن‌ها را شرح می‌دهیم. یک معماری می‌تواند

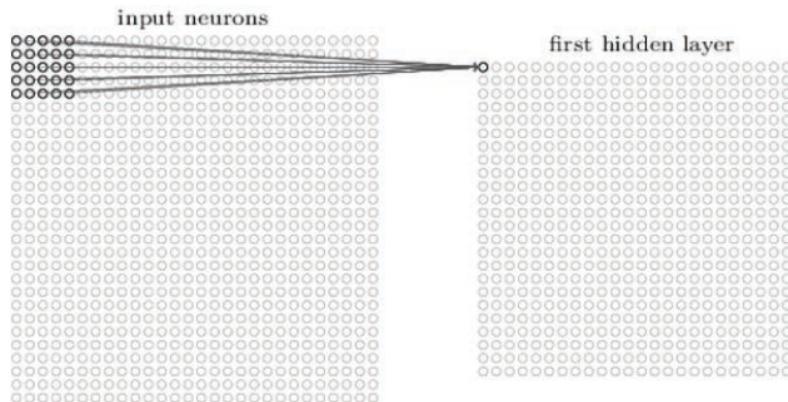
⁶³Convolutional neural network

یاد بگیرد که مسائل بسیار متفاوتی را حل کند تا زمانی که پارامترها برای هر یک از مسائل به خوبی بهینه شوند. لایه ورودی فقط نمایشی از داده خام است که به مدل داده می‌شود که نیاز به شکل ورودی ثابت دارد. در رایج ترین حالت، یک تصویر به یک آرایه 3×3^{64} ^{۶۴} تبدیل می‌شود با ابعاد $[w, h, 3]$ که w و h عرض و ارتفاع هستند. بعد آخر به دلیل استفاده از تصاویر رنگی RGB^{۶۵} اغلب ۳ است. وقتی از تصاویر اشعه ایکس^{۶۶} استفاده می‌کنیم چون دارای یک کanal^{۶۷} شدت^{۶۸} هستند بعد سوم برابر با ۱ است.

این لایه اصلی ترین لایه شبکه‌های عصبی کانولوشنی است و این شبکه‌ها نام خود را از این لایه‌ها دریافت می‌کنند. وظیفه این لایه استخراج ویژگی‌ها است. این لایه عملیات کانولوشن را بر روی داده ورودی اعمال می‌کند و خروجی‌هایی به نام نقشه ویژگی^{۶۹} از این لایه به دست می‌آید. در نتیجه تمامی نورون‌ها در یک نقشه ویژگی، وزن‌ها و بایاس‌ها^{۷۰} مشابه و مشترکی دارند که باعث می‌شود، ویژگی‌های تصویر در موقعیت‌های مختلف قابل شناسایی باشند. از طرف دیگر این اشتراک وزن‌ها باعث کاهش تعداد پارامترهای مورد نیاز برای آموزش می‌شود. در شبکه‌های کانولوشن اتصالات به صورت نواحی کوچک و محلی صورت می‌گیرد. به بیان دیگر هر نورون در نخستین لایه مخفی به ناحیه کوچکی از نورون‌های ورودی متصل می‌شود. برای مثال اگر این ناحیه 5×5 باشد این ناحیه کوچک 25×25 پیکسلی ناحیه ادراک محلی^{۷۱} یا کرنل^{۷۲} کانولوشن نامیده می‌شود. با توجه به شکل ۱۰.۲ یک تصویر ورودی 28×28 داریم که یک کرنل 5×5 بر روی پیکسل‌های ورودی از چپ به راست حرکت می‌کند هر پنجره به نورونی در لایه مخفی متصل می‌شود. بنابراین همان طور که در شکل ۱۰.۲ مشخص است لایه مخفی شامل یک شبکه 24×24 نورونی خواهد بود.

در شکل ۱۰.۲ هر نورون لایه مخفی دارای یک بایاس و تعداد 5×5 وزن می‌باشد که به ناحیه ادراکی خود متصل شده است. تمامی نورون‌های لایه مخفی مذکور که دارای ابعاد 24×24 هستند، دارای وزن‌ها و بایاس‌های مشترکی می‌باشند. به عبارت دیگر خروجی نورون لایه کانولوشن $y_{w,h,m}$ در طول و عرض w, h

⁶⁴Dimension⁶⁵Red Green Blue⁶⁶X-ray⁶⁷Channel⁶⁸Intensity⁶⁹Feature map⁷⁰Bias⁷¹Local receptive field⁷²Kernel



شکل ۱۰.۲: عملیات کانولوشن^{۷۴} در یک شبکه عصبی کانولوشنی با کرنل 5×5

عمق m به صورت رابطه ۶.۲ است.

$$y_{w,h,m} = f \left(\sum_{i=(w-1)S+1}^{(w-1)S+K} \sum_{j=(h-1)S+1}^{(h-1)S+K} \sum_{k=1}^N W_{k,m}(x_{i,j,k}) + b_m \right) \quad (6.2)$$

که در این رابطه f تابع فعالیت^{۷۵}، b_m بایاس مشترک نورون‌ها، $W_{k,m}$ وزن‌های 5×5 مشترک نورون‌ها و $x_{i,j,k}$ ورودی در موقعیت i, j, k می‌باشد. بنابراین تمامی نورون‌های واقع در لایه مخفی اول به طور دقیق ویژگی‌های مشابهی را در نواحی مختلف تصویر شناسایی می‌کنند. در نهایت خروجی لایه ورودی یا نورون‌های لایه مخفی به عنوان نقشه ویژگی شناخته می‌شوند. ابعاد مربوط به ماتریس خروجی لایه کانولوشن $\times H_2 \times W_2$ که از ماتریس ورودی با ابعاد $D_1 \times H_1 \times W_1$ است، به صورت رابطه ۷.۲ به دست می‌آید.

$$W_2 = \frac{W_1 - F + 2P}{S + 1}, \quad H_2 = \frac{H_1 - F + 2P}{S + 1}, \quad D_2 = K \quad (7.2)$$

در روابط ۷.۲ که بیانگر نحوه محاسبه ابعاد ماتریس خروجی کانولوشن است، F, P, S و k به ترتیب نشان دهنده اندازه کرنل، مدار لایه‌گذاری صفر^{۷۶}، اندازه اندازه گام^{۷۷} و تعداد فیلترها می‌باشد. طبق این روابط به ازای هر فیلتر

⁷⁵Activation function

⁷⁶Zero padding

⁷⁷Stride

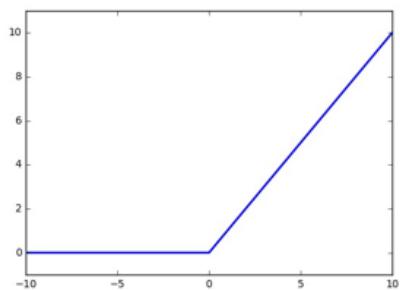
تعداد $D_1 \times F \times F \times D_1$ وزن داریم و با توجه به تعداد k فیلتر موجود، در مجموع تعداد $(F \times F \times D_1)k$ وزن k بایاس ایجاد می‌شود. بنابراین تعداد پارامترهایی که شبکه در یک لایه کانولوشن خود می‌بایست آموزش بیند زیاد است.

بکارگیری تابع فعالیت در لایه کانولوشن باعث ایجاد خصوصیات غیر خطی در خروجی می‌شود و باعث می‌شود عملکرد مدل متمایز کننده‌تر شود. این توابع با حفظ اندازه لایه، بدون نیاز به پارامترهای آموخته شده، یک عملکرد ساده عنصرگونه در مدل انجام می‌دهند. تابع تابع واحد اصلاح شده خطی^{۷۸} متداول ترین تابع مورد استفاده به خاطر آسان کردن مرحله آموزش است. مثال‌های دیگر شامل تابع سیگموید^{۷۹} و هایپربولیک^{۸۰} است.

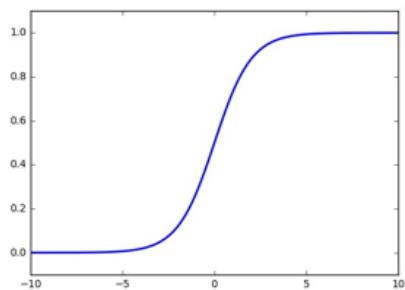
$$\text{ReLU: } r_{m,n,c} = \max\{\circ, l_{x,y,z}\}$$

$$\text{Sigmoid: } s_{m,n,c} = \frac{1}{1 - \exp(-l_{x,y,z})} \quad (8.2)$$

در یک شبکه عصبی کانولوشن معمولاً پس از هر لایه کانولوشن یک لایه pooling قرار می‌گیرد. این لایه از آن



(a) ReLU



(b) Sigmoid

شکل ۱۱.۲: (a) تابع فعالیت ReLU و (b) تابع فعالیت سیگموید

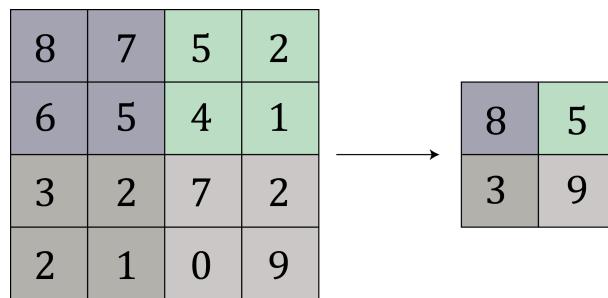
جهت اهمیت دارد که باعث کاهش تعداد پارامترهایی می‌شود که باید آموزش بینند. بنابراین با بکارگیری این لایه ضمن کاهش محاسبات مورد نیاز در بخش آموزش، باعث کنترل پیش‌پردازش^{۸۱} احتمالی در شبکه می‌شود. این لایه بر روی هر عمق از ورودی اعمال می‌شود و اندازه آن را تغییر می‌دهد. دو تابع عملکردی معروف این لایه نام دارند که تابع اول دارای کاربرد بیشتری در شبکه‌های عصبی کانولوشنی mean-pooling و max-pooling

⁷⁸Rectified linear unit (ReLU)⁷⁹Sigmoid⁸⁰Hyperbolic tangent⁸¹Over-fitting

است. طریقه عملکرد max-pooling به این صورت است که در هر پنجره بزرگترین پیکسل^{۸۲} را به خروجی می‌فرستد. این پنجره بر روی تصویر مانند تابع کانولوشن از چپ به راست و از بالا به پایین با انداه گام‌های مشخص حرکت می‌کند و نتیجه را به خروجی می‌فرستد. به دلیل اینکه این عملیات بر روی تمامی عمق‌ها اعمال می‌گردد، عمق خروجی همان عمق ورودی به لایه pooling است. یک مثال از عمل max-pooling در شکل ۱۲.۲ به نمایش گذاشته شده است.

$$\text{with } l \in [s \times x, s \times x + m], j \in [s \times y, s \times y + m], \quad R_{x,y,x} = \max\{l_{i,j,z}\} \quad (9.2)$$

لایه کاملاً متصل لایه آخر یک شبکه عصبی کانولوشنی محسوب می‌شود و اتصالات کاملی با خروجی لایه قبلی



شکل ۱۲.۲: تابع max-pooling بر روی آرایه دو بعدی کوچک $s = 2$ و $m = 2$

ایجاد می-کند. این لایه ورودی را دریافت و سپس خروجی را به صورت برداری با N مولفه تولید می‌کند که N تعداد کلاس‌هایی که شبکه باید طبقه بندی کند است. در واقع یک شبکه شبکه عصبی کانولوشنی جهت تولید یک بردار خروجی با N مولفه عددی طراحی می‌شود که هر عدد در این بردار خروجی درصد احتمال تعلق به کلاس مورد نظر را نشان می‌دهد. برای یک مسئله با تعداد k کلاس، k نورون خروجی داریم که هر احتمال را با تابع SoftMax محاسبه می‌کنند

$$P(C)_j = \frac{e^{c_j}}{\sum_{k=1}^K e^{c_k}} \quad (10.2)$$

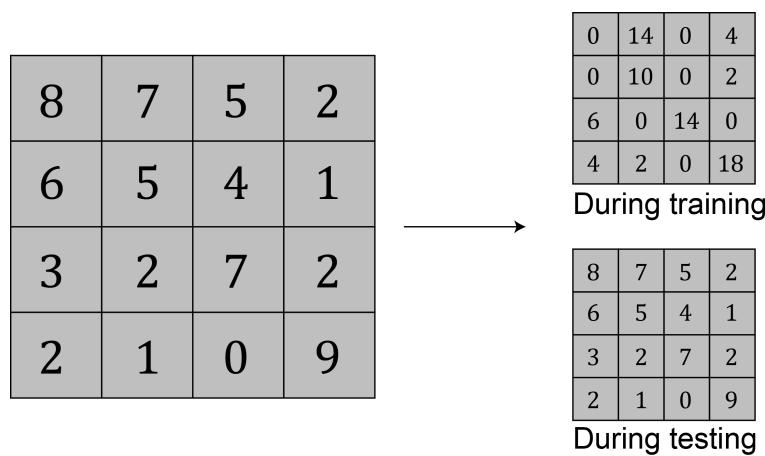
اگر دو کلاس داشته باشیم می‌توانیم از تابع SoftMax با دو خروجی استفاده کنیم یا از یک نورون استفاده کنیم و

⁸²Pixel

تابع سیگموید را محاسبه کنیم. برای دو کلاس احتمال توسط معادله σ ؟ محاسبه می‌شود

$$P(1) = \frac{1}{1 + e^{-i}} \quad P(0) = 1 - P(1) \quad (11.2)$$

حذف تصادفی^{۸۳} یک روش بسیار رایج برای جلوگیری از بیش‌پردازش شبکه عصبی مصنوعی از جمله مدل‌های یادگیری عمیق است [۷۰]. ایده این تکنیک این است که با جلوگیری از هماهنگی نورون‌ها، ویژگی‌های قوی تری ایجاد شود. اجرای آن ساده است تنها نیاز به بهم چسباندن لایه‌های اضافی در شبکه معمولاً پس از توابع فعال‌سازی است. این مازول بطور تصادفی برخی از نقاط نقشه ویژگی ورودی را صفر می‌کند. هریک از مازول‌ها دارای یک احتمال مستقل σ برای نگهداری نقاط هستند و در صورت بروز چنین اتفاقی، توسط $\frac{1}{\sigma}$ مقیاس بندی می‌شوند. نقاطی که نگهداری نمی‌شوند بر روی صفر تنظیم می‌شوند. این لایه فقط یک پارامتر σ دارد، که برای آموزش در فاصله $[0, 1]$ قرار دارد و برای آزمایش روی ۱ قرار می‌گیرد. به طور شهودی، می‌توان این فرآیند را به عنوان حذف برخی از نورون‌های شبکه عصبی، به طور موقت، همراه با اتصالات ورودی و خروجی آن تصویر کرد. مکانیزم حذف، نورون‌هایی را که به اتصالات ورودی کمتری متکی هستند را در نظر می‌گیرد. زیرا افت یک زیر مجموعه از ورودی‌ها در مقایسه با یک نورون که به بسیاری از ورودی‌ها متکی است، قابل توجه‌تر خواهد بود و به این ترتیب ویژگی‌های کلی تر مهم‌تر می‌شوند. شکل ۱۳.۲ یک مثال از لایه حذف تصادفی را نمایش می‌دهد. نرمال‌سازی دسته^{۸۴} یک تکنیک جدید ولی خیلی کارآمد است. در طی آموزش مدل‌های عمیق، وزن‌ها در هر



شکل ۱۳.۲: لایه حذف تصادفی با $\sigma = 0.5$

⁸³Dropout

تکرار^{۸۵} به روز می‌شوند. یک اثر جانبی این امر این است که در هر لایه توزیع‌های ورودی تغییر می‌کند، پدیده‌ای که به آن تغییر همبستگی داخلی^{۸۶} می‌گویند. این پدیده فرایند آموزش را کند می‌کند، به مقدار دهی دقیق‌تر وزن احتیاج دارد و مانع بهینه‌سازی^{۸۷} مدل‌های غیرخطی اشباع، مانند مماس‌های سیگموید یا هایپربولیک می‌شود. برای حل این مشکل نرم‌السازی دسته را پیشنهاد می‌شود که مشابه با حذف تصادفی، به عنوان لایه‌ای در شبکه با رفتارهای متفاوت در حین آموزش و آزمون پیاده سازی می‌شود. برای رفع مشکل تغییر کواریانس^{۸۸} داخلی، این لایه برای هر دسته آموزش با کم کردن میانگین و تقسیم بر انحراف استاندارد^{۸۹} همه نورون‌های عمق مشابه، ورودی خود را نرم‌ال می‌کند. به میانگین و انحراف استاندارد آمار mini-batch گفته می‌شود. برای اطمینان از اینکه مدل می‌تواند دقیقاً همان تابع را بیان کند، دو وزن جدید قابل تمرین^۷ و β اضافه می‌شوند که خروجی را اندازه‌گیری و جبران می‌کنند. بنابراین خروجی به صورت معادله ۱۲.۲ است.

$$\begin{aligned} \text{در طی آموزش} &: I_c = \gamma \left(\frac{I_c - \text{mean}(I_c)}{\text{std}(I_c)} \right) + \beta \\ \text{در طی آزمایش} &: I_c = \gamma \left(\frac{I_c - u_c}{v_c} \right) + \beta \end{aligned} \quad (12.2)$$

که u_c و v_c متوسط‌های در حال اجرا (I_c) و $\text{std}(I_c)$ هستند. نشان داده شده است که نرم‌السازی دسته باعث آهنگ یادگیری بالاتر می‌شود و مدل در تکرارهای کمتری همگرا خواهد شد. این روش دارای اثر رگولاrizیشن^{۹۰} است. مدل با استفاده از تابع هزینه^{۹۱} یاد می‌گیرد. این روشی است برای ارزیابی اینکه تا چه میزان خوب یک الگوریتم داده‌های مشاهده شده را می‌تواند مدل سازی کند. اگر پیش‌بینی‌ها بیش از حد از نتایج واقعی منحرف شوند، تابع هزینه مقدار بالایی خواهد داشت. به تدریج، با کمک برخی توابع بهینه سازی، تابع هزینه می‌آموزد تا خطای خطا در پیش‌بینی را کاهش دهد.

بهینه‌سازی مهمترین بخش در الگوریتم‌های یادگیری عمیق است. این کار با تعریف تابع هزینه شروع می‌شود و با به حداقل رساندن آن با استفاده از یک روش بهینه سازی به پایان می‌رسد. فرض کنید یک مجموعه داده D با تعداد I تصویر داریم. این تصاویر می‌توانند ضایعه باشند یا نباشند، بنابراین دارای برچسب $\{0, 1\} \in \mathcal{Y}$ هستند.

⁸⁴Batch normalization

⁸⁵Iteration

⁸⁶Internal covariate shift

⁸⁷Optimization

⁸⁸Covariance

⁸⁹Standard deviation

⁹⁰Regularization

⁹¹Cost function

باید مدلی بسازیم که با توجه به یک تصویر ورودی I_i , یک احتمال $(I_i)p$ تولید کند که تا حد ممکن به برچسب مربوط به آن تصویر (y_i) نزدیک باشد. برای این منظور الگوریتم‌های بهینه سازی متفاوتی وجود دارد مانند SGD^{۹۲} و Adadelta^{۹۳}.

به حداقل رساندنتابع هزینه با کاهش گرادیان تقریباً رایج ترین الگوریتم برای بهینه سازی شبکه‌های عصبی است. اگر تابع هزینه آنتروپی متقاطع دودویی^{۹۴} باشد و بخواهیم محاسبه کنیم که $(I_i)p$ تا چه حد خوب می‌تواند برچسب y_i را تقریب بزند از معادله ۱۳.۲ استفاده می‌شود.

$$L = \frac{1}{|\mathcal{D}|} \sum_i^{\mathcal{D}} \left(y_i \log(P(I_i)) + (1 - y_i) \log(1 - P(I_i)) \right) \quad (13.2)$$

احتمال برای یک ورودی به وزن‌های آن (θ) بستگی دارد و با $p(I, \theta)$ نمایش داده می‌شود. با توجه به θ می‌توان $L(\theta)$ را با اجرای مدل بر روی مجموعه داده به دست آورد.

بكپروپگیشن^{۹۴} اساس آموزش شبکه عصبی است. این عمل تنظیم-دقیق وزن‌های یک شبکه عصبی بر اساس میزان خطای^{۹۵} در هر دوره^{۹۶} قبلی است که این امر با محاسبه مشتق‌های تابع خطای بر اساس وزن‌ها $\nabla_{\theta} L(\theta)$ در زمان آموزش امکان پذیراست. تنظیم مناسب وزن‌ها باعث کاهش میزان خطای می‌شود. در فرایند بكپروپگیشن ابتدا ورودی در سراسر شبکه انتشار داده می‌شود سپس $L(\theta)$ محاسبه شده و در نهایت این خطای از طریق تمام وزن‌ها در شبکه رو به عقب منتشر می‌شود. مشتق تابع هزینه از خروجی توسط معادله ۱۴.۲ محاسبه می‌شود.

$$\frac{\partial L}{\partial P} = \frac{\partial \left(-(y_i \log(p) + (1 - y_i) \log(1 - p)) \right)}{\partial P} = \frac{P - y}{P(1 - P)} \quad (14.2)$$

همچنین محاسبه مشتق تابع هزینه L از ورودی i به صورت معادله ۱۵.۲ محاسبه می‌شود.

$$\frac{\partial L}{\partial i} = \frac{\partial L}{\partial P} \frac{\partial P}{\partial i} = P - y \quad (15.2)$$

⁹²Stochastic gradient descent

⁹³Binary cross-entropy

⁹⁴Back-propagation

⁹⁵Loss

⁹⁶epoch

همچنین محاسبه مشتق تابع هزینه بر اساس وزن‌های لایه آخر w به صورت،

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial P} \frac{\partial P}{\partial i} \frac{\partial i}{\partial w} = (P - y)a \quad (16.2)$$

می‌باشد که a در آن برابر با ترکیب خطی از ورودی‌های لایه آخر است. این کار را می‌توان به راحتی به لایه‌های قبلی تعمیم داد، بنابراین می‌توان $\nabla_{\theta} L(\theta)$ را محاسبه کرد.

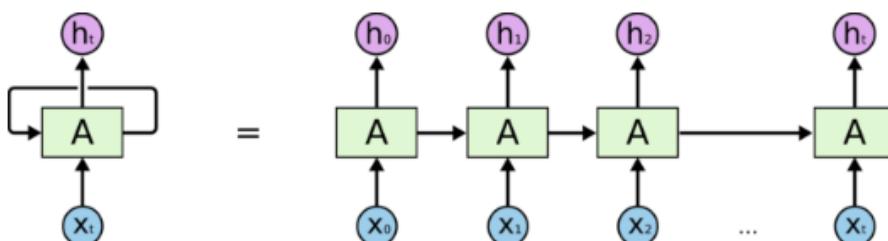
۱۱.۲ شبکه‌های عصبی بازگشتی

قبل از آشنا شدن با شبکه‌های عصبی بازگشتی^{۹۷} بهتر است مروری بر مفهوم شبکه عصبی داشته باشیم. شبکه‌های عصبی مجموعه‌ای از الگوریتم‌ها هستند که شباهت نزدیکی به مغز انسان داشته و به منظور تشخیص الگوهای طراحی شده‌اند. شبکه‌ی عصبی داده‌های حسی را از طریق ادراک ماشینی، برچسب زدن یا خوشه‌بندی ورودی‌های خام تفسیر می‌کند. شبکه می‌تواند الگوهای عددی را شناسایی کند؛ این الگوها بردارهایی هستند که همه‌ی داده‌های دنیای واقعی (تصویر، صدا، متن یا سری‌های زمانی) برای تفسیر باید به شکل آن‌ها درآیند. شبکه‌های عصبی مصنوعی از تعداد زیادی مؤلفه‌ی پردازشی (نورون) تشکیل شده‌اند که اتصالات زیادی بینشان وجود دارد و برای حل یک مسئله با یکدیگر همکاری دارند. شبکه‌ی عصبی مصنوعی معمولاً تعداد زیادی پردازشگر دارد که به صورت موازی کار می‌کنند و در ردیف‌هایی کنار هم قرار می‌گیرند. ردیف اول، همچون عصب‌های بینایی انسان در پردازش بصری، اطلاعات ورودی‌های خام را دریافت می‌کند. سپس هر کدام از ردیف‌های بعدی، به جای ورودی خام، خروجی ردیف قبلی را دریافت می‌کند؛ در پردازش بصری نیز نورون‌هایی که از عصب بینایی فاصله دارند، سیگنال را از نورون‌های نزدیک‌تر می‌گیرند. ردیف آخر خروجی کل سیستم را تولید می‌کند.

⁹⁷Recurrent Neural Network

۱۱.۲ شبکه عصبی بازگشته چیست؟

شبکه‌ی عصبی بازگشته شکلی از شبکه‌ی عصبی پیشخور است که یک حافظه‌ی داخلی دارد. شبکه عصبی بازگشته ذاتاً بازگشته است، زیرا یک تابع یکسان را برای همه‌ی داده‌های ورودی اجرا می‌کند، اما خروجی داده‌ی (ورودی) فعلی به محاسبات ورودی قبلی بستگی دارد. خروجی بعد از تولید، کپی شده و مجدداً به شبکه‌ی بازگشته فرستاده می‌شود. این شبکه برای تصمیم‌گیری، هم ورودی فعلی و هم خروجی که از ورودی قبلی آموخته شده را در نظر می‌گیرد. شبکه عصبی بازگشته شبکه‌های عصبی پیشخور می‌توانند از حالت (حافظه‌ی) درونی خود برای پردازش دنباله‌هایی از ورودی‌ها استفاده کنند. این خاصیت باعث می‌شود در مسائلی همچون تشخیص دست خط زنجیره‌ای یا تشخیص گفتار کاربرد داشته باشند. در سایر شبکه‌های عصبی، ورودی‌ها از یکدیگر مستقل هستند، اما در شبکه عصبی بازگشته ورودی‌ها به هم مرتبط می‌باشند. به شکل ۱۴.۲ توجه کنید، این شبکه ابتدا X_0 را از دنباله‌ی ورودی‌ها گرفته و خروجی h_0 را تولید می‌کند که همراه با X_1 ورودی گام بعدی



An unrolled recurrent neural network.

شکل ۱۴.۲: یک نمونه بازشده شبکه عصبی بازگشته

محسوب خواهند شد. یعنی X_1 ورودی گام بعدی هستند. به همین صورت h_1 بعدی همراه با X_1 ورودی گام بعدی خواهند بود. شبکه عصبی بازگشته بدین طریق می‌تواند هنگام آموزش زمینه را به خاطر داشته باشد. فرمول حالت^{۹۸} کنونی به صورت رابطه ۱۷.۲ خواهد بود که در آن،

$$h_t = f(h_{t-1}, x_t) \quad (17.2)$$

⁹⁸State

خواهد بود که در آن h_t برابر است با،

$$h_t = \tanh(W_{hh}h_{t-1} + W_{hx}x_t) \quad (18.2)$$

در این فرمول W وزن، h تکبردار نهان، $W_h h$ وزن حالت نهان قبلی، W_{hx} وزن حالت ورودی کنونی و \tanh تابع فعالیت است که با استفاده از تابعی غیرخطی، خروجی را فشرده می‌کند تا در بازهی $[1, -1]$ جای گیرند. در نهایت حالت خروجی y_t از طریق رابطه ۱۹.۲ بدست می‌آید،

$$y_t = W_{hy}h_t \quad (19.2)$$

که در آن W_{hy} برابر وزن در حالت تولید شده را نشان می‌دهد.

۲.۱۱.۲ مزایای شبکه عصبی بازگشتی

شبکه عصبی بازگشتی می‌تواند دنباله‌ای از داده‌ها را به شکلی مدل‌سازی کند که هر نمونه وابسته به نمونه‌های قبلی به نظر برسد. شبکه عصبی بازگشتی را می‌توان با لایه‌های پیچشی نیز به کار برد تا گستره‌ی همسایگی پیکسلی را افزایش داد.

۳.۱۱.۲ معایب شبکه عصبی بازگشتی

- گرادیان کاهشی و مشکلات ناشی از آن
- آموزش بسیار دشوار
- ناتوانی در پردازش دنباله‌های طولانی از ورودی در صورت استفاده از تابع فعالیت \tanh یا $ReLU$

۴.۱۱.۲ کاربردهای شبکه عصبی بازگشته

- شرح نویسی عکس^{۹۹}: شبکه عصبی بازگشته با تحلیل حالت کنونی عکس، برای شرح نویسی عکس به کار می‌رود
- پیش‌بینی سری‌های زمانی^{۱۰۰}: هر مسئله سری زمانی مانند پیش‌بینی قیمت یک سهام در یک ماه خاص، با شبکه عصبی بازگشته قابل انجام است
- پردازش زبان طبیعی^{۱۰۱}: کاوش متن و تحلیل احساسات می‌تواند با استفاده از شبکه عصبی بازگشته انجام شود
- ترجمه ماشینی^{۱۰۲}: شبکه شبکه عصبی بازگشته می‌تواند ورودی خود را از یک زبان دریافت و آن را به عنوان خروجی به زبان دیگری ترجمه کند

۵.۱۱.۲ انواع شبکه عصبی بازگشته

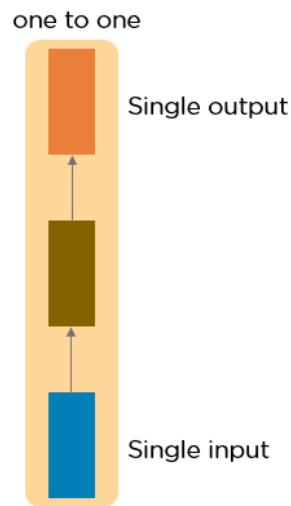
- به طور کلی ۴ نوع شبکه عصبی بازگشته داریم:
- یک به یک (one to one): این نوع شبکه عصبی به عنوان شبکه عصبی وانیلی نیز شناخته می‌شود و برای مسائل یادگیری ماشین که یک ورودی و یک خروجی دارند به کار می‌رود.
 - یک به چند (one to many): این شبکه عصبی بازگشته دارای یک ورودی و چند خروجی است. یک نمونه آن، شرح نویسی عکس است.
 - چند به یک (many to one): این نوع از شبکه عصبی بازگشته، دنباله ایی از ورودی‌ها را می‌گیرد و یک خروجی تولید می‌کند. تحلیل احساسات مثال خوبی از این نوع شبکه است که یک جمله را به عنوان ورودی می‌گیرد و آن را با احساس مثبت یا منفی طبقه بندی می‌کند.
 - چند به چند (many to many): دنباله ایی از ورودی‌ها را می‌گیرد و دنباله ایی از خروجی‌ها را تولید می‌کند. ترجمه ماشینی نمونه ایی از این نوع شبکه است.

⁹⁹Image Captioning

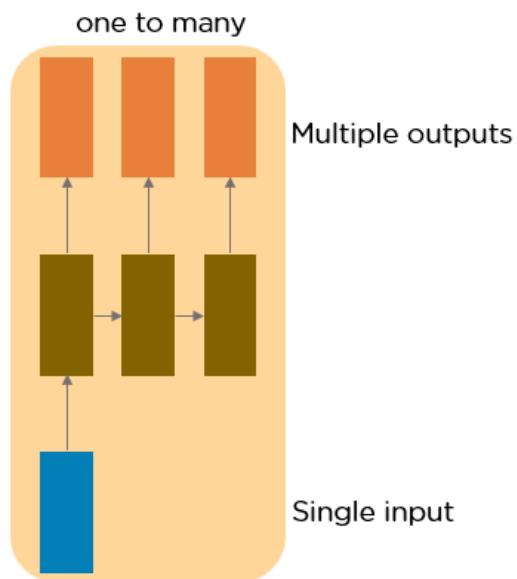
¹⁰⁰Time Series Prediction

¹⁰¹Natural Language Processing

¹⁰²Machine Translation



شکل ۱۵.۲: ساختار شبکه عصبی بازگشتی یک به یک

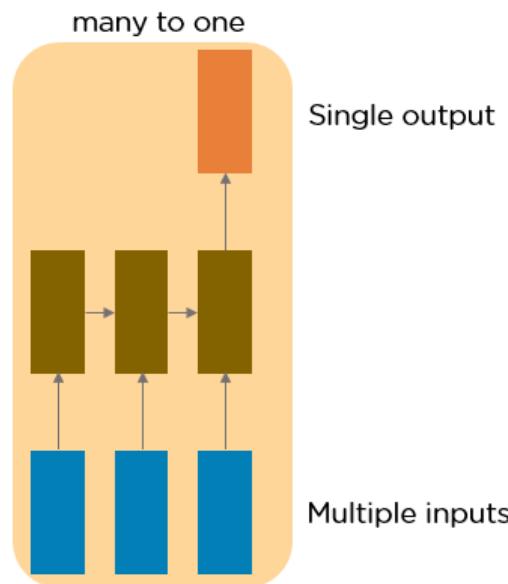


شکل ۱۶.۲: ساختار شبکه عصبی بازگشتی یک به چند

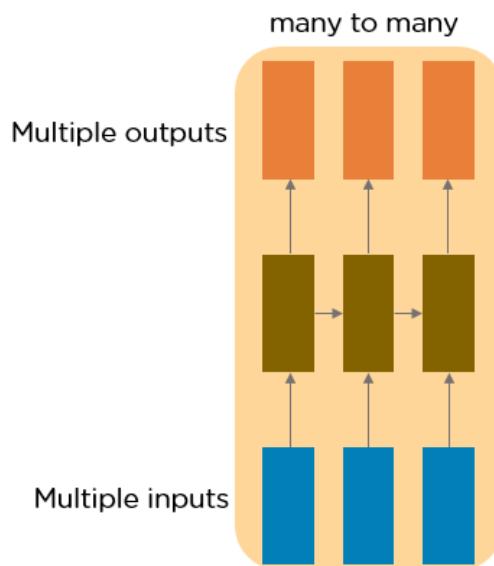
۶.۱۱.۲ حافظه‌ی کوتاه‌مدت بلند (LSTM)

شبکه‌های حافظه‌ی کوتاه‌مدت بلند^{۱۰۳} یا LSTM نسخه‌ی تغییریافته‌ای از شبکه‌های عصبی بازگشتی هستند که یادآوری داده‌های گذشته در آن‌ها تسهیل شده است. مشکل گرادیان کاهشی که در شبکه عصبی بازگشتی وجود

¹⁰³Long Short Term Memory (LSTM)



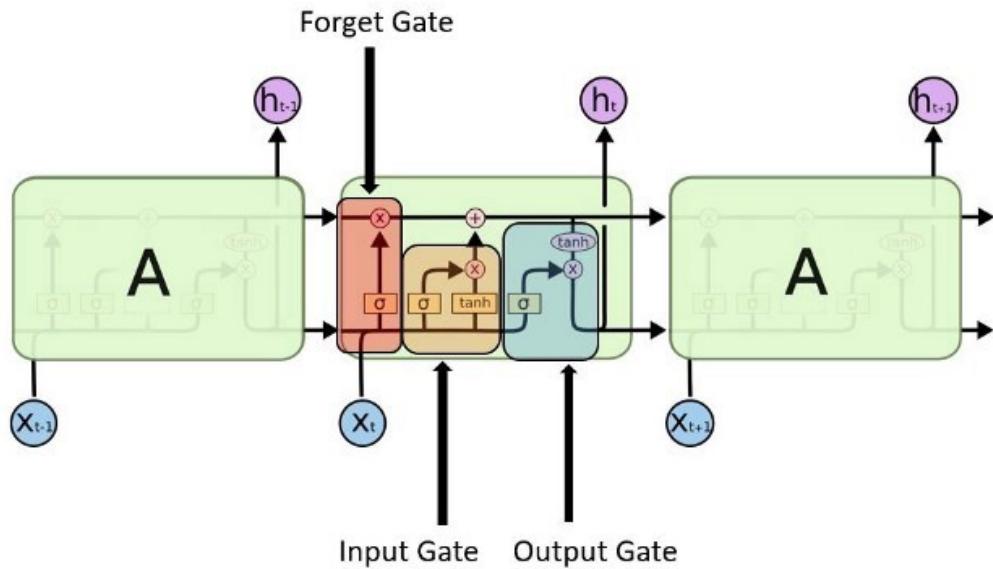
شکل ۱۷.۲: ساختار شبکه عصبی بازگشتی چند به یک



شکل ۱۸.۲: ساختار شبکه عصبی بازگشتی چند به چند

داشت نیز در این شبکه‌ها حل شده است. شبکه‌های LSTM برای مسائل رده‌بندی، پردازش و پیش‌بینی سری‌های زمانی با استفاده از برجسب‌های زمانی مدت‌های نامعلوم مناسب هستند. این شبکه‌ها مدل را با استفاده از انتشار رو به عقب آموزش می‌دهند.

همان‌طور که در شکل ۱۹.۲ نمایش داده شده است، در یک شبکه‌ی LSTM سه دریچه وجود دارد:



شکل ۱۹.۲: ساختار LSTM

دریچه‌های LSTM

۱) **دریچه‌ی ورودی:** با استفاده از این دریچه می‌توان دریافت کدام مقدار از ورودی را باید برای تغییر حافظه به کار برد. تابع سیگموید تصمیم می‌گیرد مقادیر بین ۰ و ۱ اجازه‌ی ورود دارند و تابع \tanh با ضریب‌دهی (بین ۱ تا ۱+) به مقادیر، در مورد اهمیت آن‌ها تصمیم می‌گیرد.

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \end{aligned} \quad (۲۰.۲)$$

۲) **دریچه‌ی فراموشی:** از طریق این دریچه می‌توان جزئیاتی را که باید از بلوک حذف شوند، تشخیص داد. تصمیم‌گیری در این مورد بر عهده‌ی تابع سیگموید است. این تابع با توجه به حالت قبلی h_{t-1} و ورودی محظوظ X_t ، عددی بین ۰ تا ۱ به هر کدام از اعداد موجود در حالت سلولی C_{t-1} اختصاص می‌دهد؛ نشان‌دهنده‌ی حذف

آن عدد و ۱ به معنی نگه داشتن آن است.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (21.2)$$

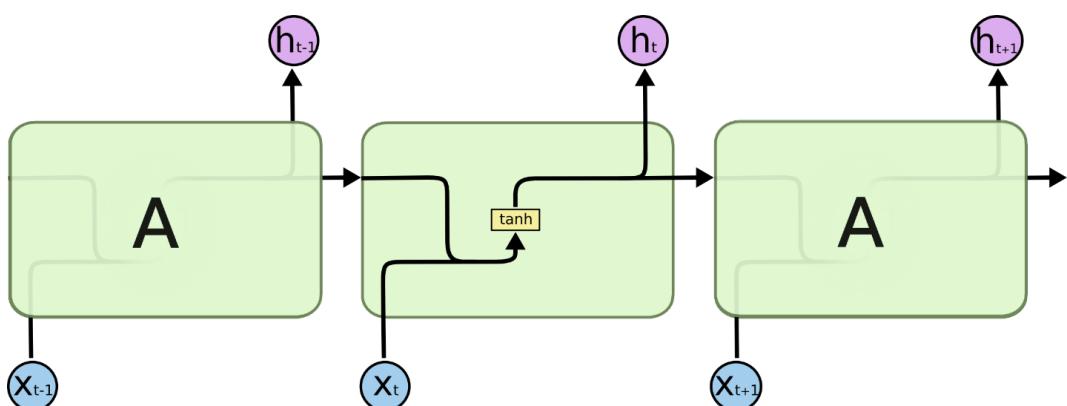
(۳) دریچه خروجی: ورودی و حافظه بلوک برای تصمیم‌گیری در مورد خروجی مورد استفاده قرار می‌گیرند.تابع سیگموئید تصمیم می‌گیرد مقادیر بین ۰ و ۱ اجازه بودن دارند و تابع \tanh با ضریب دهی (بین ۱ تا +۱) به مقادیر و ضرب آنها در خروجی تابع سیگموید در مورد اهمیت آنها تصمیم‌گیری می‌کند.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (22.2)$$

$$h_t = o_t * \tanh(C_t)$$

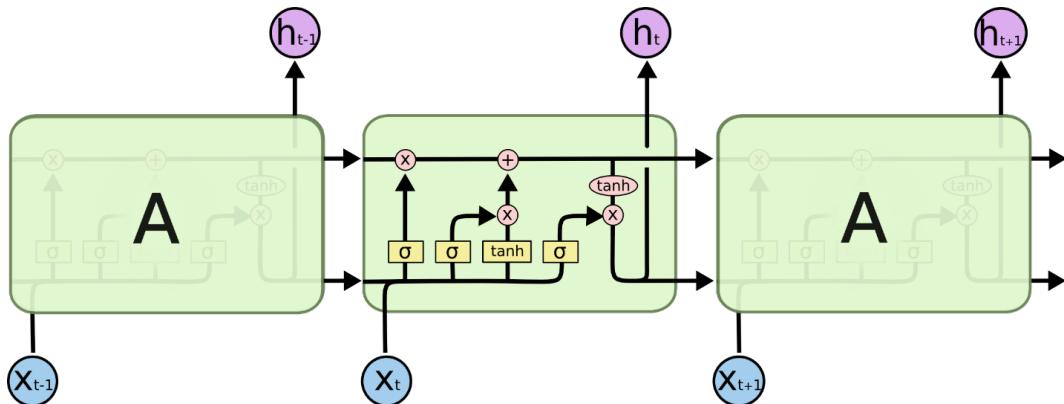
در حقیقت هدف از طراحی شبکه‌های LSTM، حل کردن مشکل وابستگی بلندمدت بود. به این نکته مهم توجه کنید که به یاد سپاری اطلاعات برای بازه‌های زمانی بلند مدت، رفتار پیش‌فرض و عادی شبکه‌های LSTM است و ساختار آنها به صورتی است که اطلاعات خیلی دور را به خوبی یاد می‌گیرند که این ویژگی در ساختار آنها نهفته است.

همه شبکه‌های عصبی بازگشتی به شکل دنباله‌ای (زنجره‌ای) تکرار شونده از مازول‌های (واحدهای) شبکه‌های عصبی هستند. در شبکه‌های عصبی بازگشتی استاندارد، این مازول‌های تکرار شونده ساختار ساده‌ای دارند، برای مثال تنها شامل یک لایه تائزانتِ هایپربولیک (\tanh) هستند.



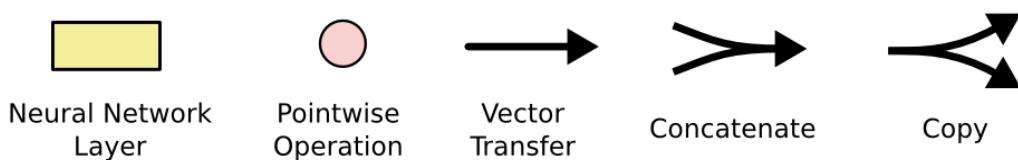
شکل ۲۰.۲: مازول‌های تکرار شونده در شبکه‌های عصبی بازگشتی استاندارد فقط دارای یک لایه هستند.

شبکه‌های LSTM نیز چنین ساختار دنباله یا زنجیره‌مانندی دارند ولی مازول تکرار شونده ساختار متفاوتی دارد. به جای داشتن تنها یک لایه شبکه عصبی، ۴ لایه دارند که طبق ساختار ویژه‌ای با یکدیگر در تعامل و ارتباط هستند. در ادامه قدم به قدم ساختار شبکه‌های حافظه‌ی کوتاه‌مدت بلند را توضیح خواهیم داد. اما در ابتدا معنی هستند.



شکل ۲۱.۲: مازول‌های تکرار شونده در LSTM‌ها دارای ۴ لایه هستند که با هم در تعامل می‌باشند.

هر کدام از شکل و علامت‌هایی را که از آن‌ها استفاده خواهیم کرد توضیح می‌دهیم. در شکل ۲۲.۲، هر خط



شکل ۲۲.۲: اشکال از راست به چپ به ترتیب برابر هستند با: کپی کردن، وصل کردن، بردار انتقال، عملیات نقطه به نقطه، یک لایه‌ی شبکه عصبی.

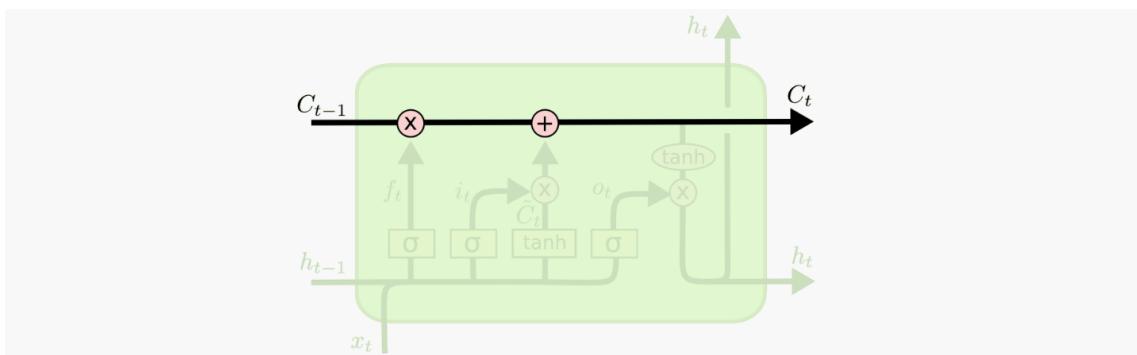
یک بردار را به صورت کامل از خروجی یک گره به ورودی گره دیگر انتقال می‌دهد. دایره‌های صورتی نمایش دهنده عملیات‌های نقطه به نقطه مانند «جمع کردن دو بردار» هستند. مستطیل‌های زرد، لایه‌های شبکه‌های عصبی هستند که شبکه پارامترهای آن‌ها را یاد می‌گیرد. خط‌هایی که با هم ادغام می‌شوند نشان‌دهنده الحاق^{۱۰۴} و خط‌هایی که چند شاخه می‌شوند نشان‌دهنده‌ای این موضوع است که محتوای آن‌ها کپی و به بخش‌های مختلف ارسال می‌شود.

عنصر اصلی LSTM‌ها سلول حالت^{۱۰۵} است که در حقیقت یک خط افقی است که در بالای شکل ۲۳.۲ قرار

¹⁰⁴Concatenation

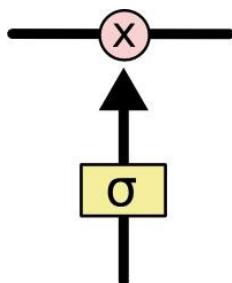
¹⁰⁵Cell state

دارد. سلول حالت را می‌توان به صورت یک تسمه نقاله تصور کرد که از اول تا آخر دنباله یا همان زنجیره با تعاملات خطی جزئی در حرکت است (یعنی ساختار آن بسیار ساده است و تغییرات کمی در آن اتفاق می‌افتد).



شکل ۲۳.۲: سلول حالت در مازول LSTM

LSTM این توانائی را دارد که اطلاعات جدیدی را به سلول حالت اضافه یا اطلاعات آن را حذف کنید. این کار توسط ساختارهای دقیقی به نام دروازه‌ها^{۱۰۶} انجام می‌شود. دروازه‌ها راهی هستند برای ورود اختیاری اطلاعات. آن‌ها از یک لایه شبکه عصبی سیگموید به همراه یک عملگر ضرب نقطه به نقطه تشکیل شده‌اند.



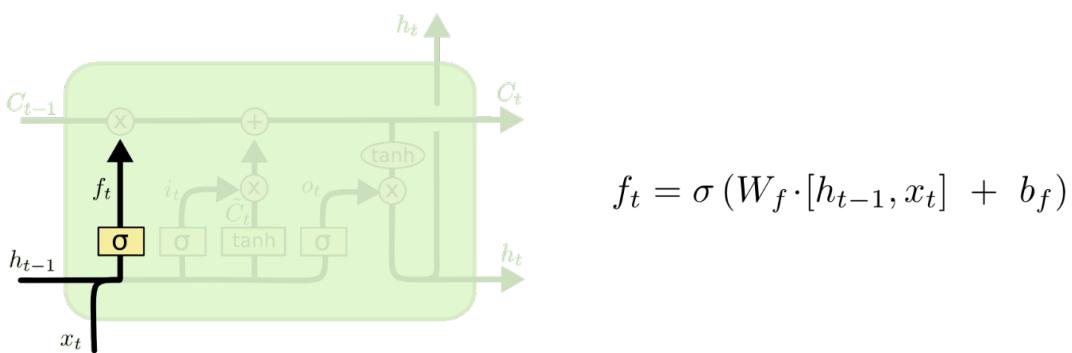
شکل ۲۴.۲: نمایی از نحوه تاثیر و ورود اطلاعات به سلول حالت

خروجی لایه سیگموید عددی بین صفر و یک است، که نشان می‌دهد چه مقدار از ورودی باید به خروجی ارسال شود. مقدار صفر یعنی هیچ اطلاعاتی نباید به خروجی ارسال شود در حالی که مقدار یک یعنی تمام ورودی به خروجی ارسال شود!

LSTM دارای ۳ دروازه مشابه برای کنترل مقدار سلول حالت است که در ادامه به بررسی قدم به قدم آن‌ها از لحظه ورود تا خروج اطلاعات خواهیم پرداخت.

¹⁰⁶Gate

قدم اول در LSTM تصمیم در مورد اطلاعاتی است که می‌خواهیم آن‌ها را از سلول حالت پاک کنیم. این تصمیم توسط یک لایه سیگموید به نام «دروازه فراموشی^{۱۰۴}» انجام می‌شود. این دروازه با توجه به مقادیر x_t و h_{t-1} برای هر عدد، مقدار صفر یا یک را در سلول حالت C_{t-1} به خروجی می‌برد. مقدار یک یعنی به صورت کامل مقدار حال حاضر سلول حالت C_{t-1} را به C_t انتقال داده شود و مقدار صفر یعنی به صورت کامل اطلاعات سلول حالت کنونی C_{t-1} را پاک شود و هیچ مقداری از آن به C_t برد نشود. بباید به مثال قبلی مان که یک مدل زبانی‌ای بود که در آن تلاش داشتیم کلمه بعدی را بر اساس همه کلمه‌های قبلی حدس بزنیم، برگردیم. در چنین مسأله‌ای، سلول حالت ممکن است در بردارنده جنسیت فاعل کنونی باشد، که با توجه به آن می‌توانیم تشخیص دهیم از چه ضمیری باید استفاده کنیم. زمانی که یک فاعل جدید در جمله ظاهر می‌شود، می‌بایست جنسیت فاعل قبلی حذف شود.



شکل ۲۵.۲: قدم اول در پاک کردن اطلاعات از سلول حالت در وضعیت ورودی

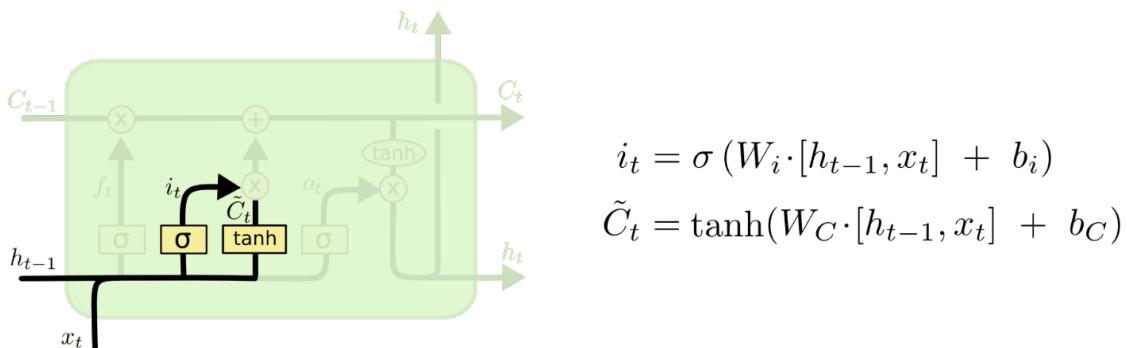
قدم بعدی این است که تصمیم بگیریم چه اطلاعات جدیدی را می‌خواهیم در سلول حالت ذخیره کنیم. این تصمیم دو بخشی است. ابتدا یک لایه سیگموید به نام دروازه ورودی^{۱۰۵} داریم که تصمیم می‌گیرد چه مقادیری به روز خواهند شد. مرحله بعدی یک لایه تائزانت هایپربولیک است که برداری از مقادیر به نام \tilde{C}_t می‌سازد که می‌توان آن‌ها را به سلول حالت اضافه کرد. در مرحله بعد، ما این دو مرحله را با هم ترکیب می‌کنیم تا مقدار سلول حالت را به روز کنیم.

در مثال مدل زبانی‌ای که پیش‌تر داشتیم، قصد داریم جنسیت فاعل جدید را به سلول حالت اضافه کنیم تا جایگزین جنسیت فاعل قبلی شود که در مرحله قبلی تصمیم گرفتیم آن را فراموش کنیم.

حال زمان آن فرا رسیده است که سلول حالت قدیمی یعنی C_{t-1} را سلول حالت جدید یعنی C_t به روز کنیم. در

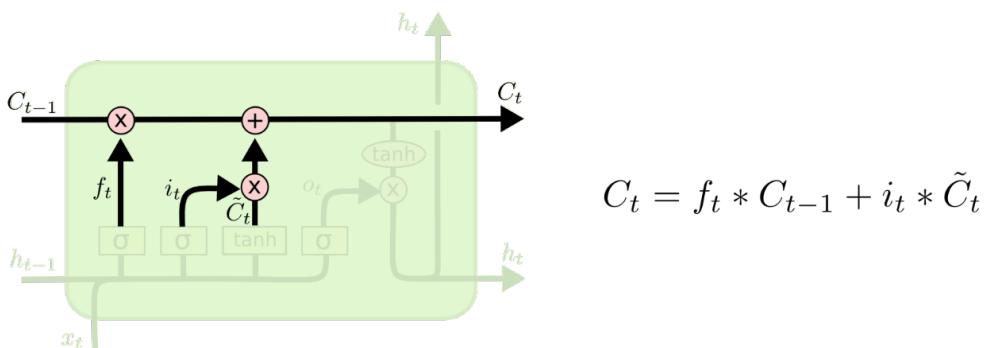
¹⁰⁷Forget gate

¹⁰⁸Input gate



شکل ۲۶.۲: قدم دوم در اضافه کردن اطلاعات جدید به سلول حالت

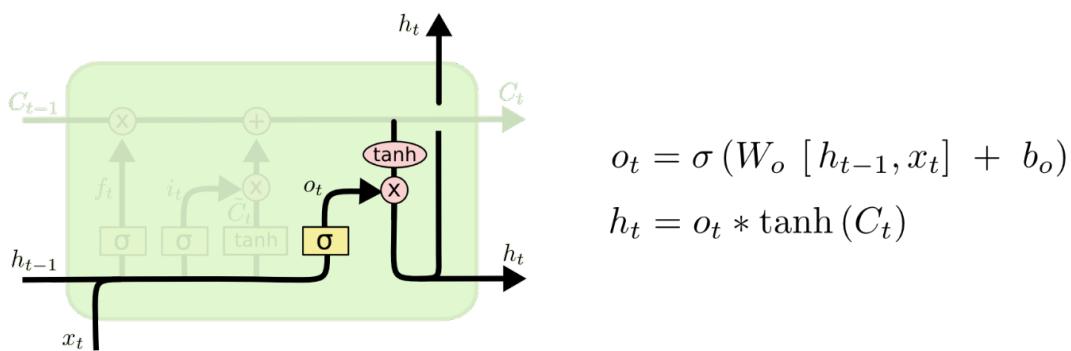
مراحل قبلی تصمیم گرفته شد که چه کنیم و در حال حاضر تنها لازم است تصمیماتی را که گرفته شد عملی کنیم. ما مقدار قبلی سلول حالت را در f_t ضرب می کنیم که یعنی فراموش کردن اطلاعاتی که پیشتر تصمیم گرفتیم آنها را فراموش کنیم. سپس $i_t * \tilde{C}_t$ را به آن اضافه می کنیم. در حال حاضر مقادیر جدید سلول حالت با توجه به تصمیماتی که پیشتر گرفته شده بود بدست آمده اند. در مثال مدل زبانی، اینجا دقیقاً جائی است که اطلاعاتی که در مورد جنسیت قبلی داشتم را دور می ریزیم و اطلاعات جدید را اضافه می کنیم.



شکل ۲۷.۲: بهروزرسانی اطلاعات در سلول حالت

در نهایت باید تصمیم بگیریم قرار است چه اطلاعاتی را به خروجی ببریم. این خروجی با در نظر گرفتن مقدار سلول حالت خواهد بود، ولی از فیلتر مشخصی عبور خواهد کرد. در ابتدا، یک لایه سیگموید داریم که تصمیم می گیرد چه بخشی از سلول حالت قرار است به خروجی برده شود. سپس مقدار سلول حالت (پس از بهروز شدن در مراحل قبلی) را به یک لایه تائزانت هایپربولیک (تا مقادیر بین -1 و $+1$ باشند) می دهیم و مقدار آن را در خروجی لایه سیگموید قبلی ضرب می کنیم تا تنها بخش هایی که مد نظرمان است به خروجی برود. در مثال مدل زبانی، با توجه به اینکه تنها فاعل را دیده است، در صورتی که بخواهیم کلمه بعدی را حدس بزنیم،

ممکن است بخواهد اطلاعاتی در ارتباط با فعل را به خروجی ببرد. برای مثال ممکن است اینکه فاعل مفرد یا جمع است را به خروجی ببرد، که ما با توجه به آن بدانیم فعل به چه فرمی خواهد بود.



شکل ۲۸.۲: قدم نهایی برای تولید خروجی ماذول LSTM

۱۲.۲ یادگیری تقویتی

۱.۱۲.۲ مقدمه و بیشینه تاریخی

ادوارد ثورندایک^{۱۰۹} پدر روانشناسی مدرن در سال ۱۸۷۴ میلادی در ایالت ماساچوست آمریکا متولد شد. اوی در اوایل قرن ۲۰ میلادی آزمایشی انجام داد که باعث ارائه قانون اثر شد. او برای این آزمایش، گربه‌ای را در جعبه‌ای موسوم به جعبه معمرا قرار داد. هر کوشش درستی، از این گربه برای نجات از جعبه صورت می‌گرفت، باعث می‌شد ثورندایک به عنوان پاداش به او غذا بدهد. به تدریج گربه به کارهای درست خود پی برد و آنها را تکرار کرد، تا جایی که دیگر هیچ کار اشتباهی نمی‌کرد و بالاخره موفق به خروج از جعبه شد. ثورندایک در سال ۱۹۱۲ به ریاست انجمن روانشناسان، در سال ۱۹۱۷ به عضویت انجمن علوم، در سال ۱۹۳۴ به ریاست انجمن علوم پیشرفته نایل آمد و در سال ۱۹۴۷ در سن ۷۴ سالگی، بدرود حیات گفت. در سال ۲۰۰۲ رتبه‌ای از برترین روانشناسان تاریخ ارائه شد که ثورندایک جزو ۱۰ روانشناس برتر تاریخ قرار گرفت. می‌توان مهم‌ترین کشف‌وی را، اثبات وجود یادگیری تقویتی در روانشناسی دانست.

شاید ریچارد بلمن^{۱۱۰} (مخترع الگوریتم بلمن-فورد) را بتوان اولین کسی دانست که یادگیری تقویتی را وارد

¹⁰⁹Edward Thorndike

¹¹⁰Richard E. Bellman

هوش مصنوعی ساخت. در اوایل دهه ۱۹۵۰ بلمن مسئله‌ای با عنوان «کنترل بهینه» را مطرح ساخت که با استفاده از روش‌های پویا در برنامه ریزی پویا کنترل کننده‌ها را به سمت نتیجه بهینه رهنمون می‌شد. در اواخر دهه ۵۰ میلادی مینسکی در پایان نامه دکتری خود روش‌های محاسبات آزمون و خطا توسط مفهوم یادگیری تقویتی را مطرح نمود و الگوریتم‌های یادگیری تقویتی را پایه ریزی کرد. در کل دهه ۵۰ میلادی را میتوان دهه تشکیل الگوریتم‌های محاسباتی اولیه یادگیری تقویتی دانست. در دهه ۶۰ میلادی اولین کابرد‌های یادگیری تقویتی به وقوع پیوستند. در اولین تلاش‌ها فارلی و کلارک از یادگیری تقویتی برای تشخیص الگو استفاده کردند بدین صورت که هر بار برنامه نتیجه بهتری به دست می‌آمد او را تشویق می‌کردند. در اواخر دهه ۶۰ میلادی، یادگیری نظارتی از یادگیری تقویتی، مشتق شد. در یادگیری نظارتی طراح نتیجه نهایی را در دست دارد و از هوش مصنوعی می‌خواهد هر بار مسیر بین ورودی و نتیجه را طراحی کرده و هر بار که برنامه، مسیر بهتری به دست می‌آورد، تشویق می‌شود. همچنین طراح نظارت مستقیم بر عملکرد عامل دارد.

فصل ۳

روش‌های پیشین

۱.۳ مقدمه

در فصل گذشته به معرفی مفاهیم و موضوعات مرتبط با این حوزه پرداخته شد. در ادامه در این فصل با توجه به اطلاعاتی که کسب کرده‌اید به معرفی و بررسی روش‌هایی که مرتبط با موضوع این پایان‌نامه است پرداخته خواهد شد و نتایج آن‌ها را برای فرض‌های و داده‌های ورودی خود مشاهده خواهیم نمود. در این بین تا جایی که ممکن باشد به بررسی نقاط قوت و ضعف آن‌ها نیز خواهیم پرداخت و در انتهای این فصل یک جدول مقایسه بین روش‌هایی که تا به حال معرفی شده‌اند را ارائه خواهیم داد.

۲.۳ مدل کیم و سایمون [۴۵]

این مدل در سال ۲۰۱۴ با تمرکز بر ساخت درخت فیلوجنی^۱ از طریق رابطه ترکیبی میان جهش‌های ایجاد شده در داده‌های توالی‌یابی تکسلولی دی‌ان‌ای ارائه گردید. بررسی رابطه‌ی ترتیبی هر یک از جهش‌های رخ داده با یکدیگر، این امکان را فراهم می‌آورد تا اطلاعاتی در مورد نحوه تشکیل کلون‌ها و ترتیب زمانی رخ دادن جهش‌های گوناگون بدست آید. همچنین امکان محاسبه نسبت زمانی سپری شده میان جهش‌های اولیه موجود در داده‌های توالی‌یابی تکسلولی تا نزدیک‌ترین جد مشترک وجود دارد. استنباط درخت فیلوجنی از طریق لگوریتم کیم و

¹Phylogeny tree

جدول ۱.۳: مثالی از چند نمونه با بررسی وجود یا عدم وجود دو جهش X و Y .

Sample	1	2	3	4	5	6	7
X mutation	0	0	0	0	1	1	0
Y mutation	0	0	1	1	1	1	1

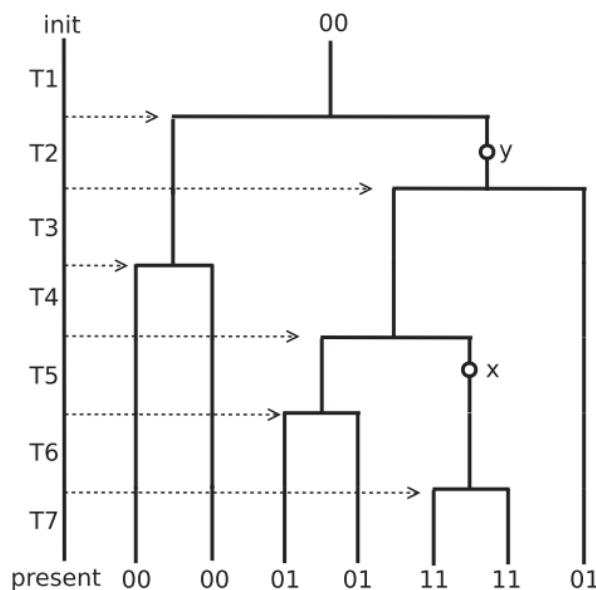
سایمون، بر مبنای منطق بیزی است، یعنی از این منطق به منظور تعیین رابطه ترتیبی بین هر دو جهش گوناگون استفاده شده است. در ادامه مقدار بیشینه درست‌نمایی^۲ درخت استباط شده بر مبنای احتمال ترتیبی دو به دوی بین هر دو جهش مختلف در دو جایگاه از یک دنباله، که از طریق ژنولوژی متفاوت با جهش‌های گوناگون در گره‌های درخت به هم مرتبط می‌شوند، محاسبه می‌شود. سرانجام مقادیر بیشینه‌ی احتمالات با شرط کمینه کردن میزان تفاوت با داده‌های مشاهده شده محاسبه می‌گردد.

از نکات قوت این الگوریتم در نظر گرفتن خطای توالی‌بایی و ترک آلل^۳ است. این عدم قطعیت در داده‌ها از طریق محاسبه بیشینه درست‌نمایی ترتیبی هر یک از جهش‌ها بدست خواهد آمد. به عنوان مثال در نظر بگیرید که هفت زوج مرتب از جهش‌های یک دی‌ان‌ای موجود است. برای سادگی بیشتر مولفه اول را با x و مولفه دوم را با y نشان داده می‌شود. داده‌های نمونه‌گیری شده از این دی‌ان‌ای در جدول ۱.۳ نشان داده شده است. در این جدول صفر بیانگر عدم وجود جهش و یک بیانگر وجود جهش است. تعداد رخداد جهش‌ها با فرض عدم وجود خطأ در توالی‌بایی داده‌ها، برابر یک در نظر گرفته می‌شود، یعنی در هر موقعیت تنها یکبار جهش رخ داده است. همچنین ترتیب زمانی رخداد جهش‌ها یک ترتیب جزئی است، به این معنی که زوج (۱،۱) بیانگر این است که یا جهش x مقدم بوده است یا جهش y . زوج (۱،۰) بیانگر آن است که جهش x وجود نداشته است ولی جهش y وجود داشته و با فرض اینکه هیچ جهشی از بین نمی‌رود، در نتیجه می‌توان استباط کرد که y نسبت به x قدیمی‌تر است و به عنوان یکی از اجداد x در درخت فیلوزنی تومور قرار می‌گیرد. در نتیجه با استفاده از جدول داده‌های نمونه‌برداری شده، استباط یک رابطه زمانی میان جهش‌های صورت گرفته امکان پذیر است.

شکل ۱.۳ یک درخت فیلوزنیک تومور را نشان می‌دهد که از داده‌های جدول بالا استباط شده است. در این همه هفت نمونه به عنوان برگ‌های درخت مشاهده می‌شود و ریشه درخت زوج (۰،۰) می‌باشد به این معنی که در ابتدا هیچ جهشی رخ نداده است. محور عمودی بیانگر سیر زمانی تکامل تومور است که به تعداد نمونه‌ها تقسیم شده

²Maximum-likelihood³Allele dropout

است. برای استنباط درخت فیلوزنی تومور، الگوریتم کیم و سایمون از سه بخش اصلی تشکیل شده است. طبق



شکل ۱.۳: نمایی از یک درخت فیلوزنیک تومور

قضیه بیز برای محاسبه هر یک از این سه احتمال به مقادیر درست‌نمایی^۴ نیاز داریم. مقدار احتمال رخداد طبق رابطه زیر محاسبه می‌گردد:

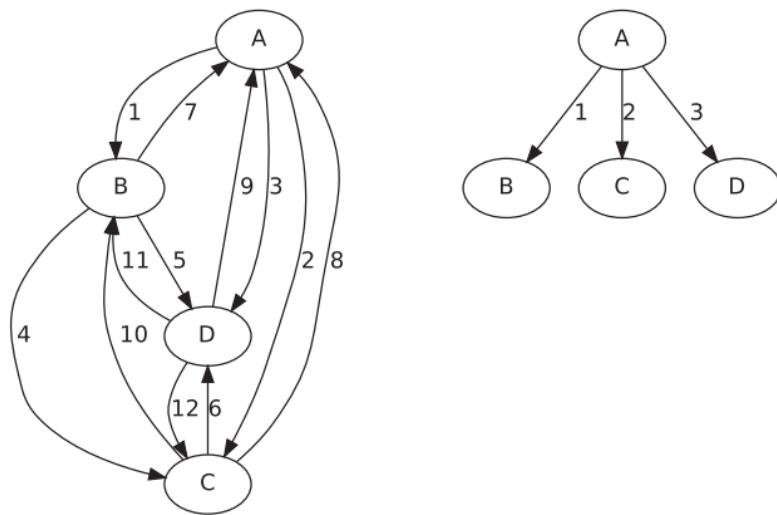
$$P(x \sim y | D) \propto L(x \sim y) P(x \sim y), \quad L(x \sim y) = P(D | x \sim y) \quad (1.3)$$

طبق این رابطه و با توجه به اینکه رابطه زمانی میان جهش‌های x و y دارای ۳ حالت،

$$x \rightarrow y, \quad y \rightarrow x \quad \text{و} \quad x \not\rightarrow y$$

است، مقدار احتمال محاسبه شده از رابطه فوق به ازای یکی از این سه حالت بیشینه است و به ازای آن حالت یک مسیر جهت‌دار در درخت فیلوزنی قرار خواهد گرفت. طبق آنچه گفته شد یک گراف جهت دار فیلوزنی بلقوه مشابه آنچه در شکل ۲.۳ نشان داده شده است استنباط خواهد شد. در نهایت از این گراف جهت دار، یک درخت به طوری که روابط میان جهش‌ها از آن استنباط شود ساخته خواهد شد. در ابتدا یال‌های گراف از طریق رابطه‌ای

⁴Likelihood



شکل ۲.۳: یک گراف جهت دار فیلوزنی

که در ادامه آمده است وزن‌دهی می‌شوند،

$$w_{x \sim y} = -\log P(x \sim y | D) \quad (2.3)$$

که در آن $(x \sim y)$ رابطه بین جهش‌های x و y است و D نمونه یا سمپل‌های موجود در داده است. بهترین درخت \hat{T} از طریق کمینه‌کردن وزن‌های گراف بدست می‌آید.

$$\hat{T} = \arg \min \left(\sum_{x \sim y \in T} w_{x \sim y} \right) = \arg \max \left(\prod_{x \sim y \in T} P(x \sim y | D) \right) \quad (3.3)$$

در شکل ۲.۳ متحتمل‌ترین درخت فیلوزنی با بیشینه درست‌نمایی بر اساس اطلاعات نمونه‌برداری شده بدست می‌آید. در این شکل گراف اولیه و درخت متناظر آن مشاهده می‌شود. مجموع همه وزن‌ها در درخت نهایی با شرط کمینه‌سازی برابر هفت است که این مقدار کمترین مقدار ممکن است.

۱.۲.۳ پایگاه داده

در این مقاله از پایگاه داده تولی‌یابی تک‌سلولی هو و همکاران [۴۱] استفاده شده است. این مجموع داده از توالی‌یابی تک‌سلولی دی‌ان‌ای نمونه‌های توموری یک نوع خاص از سرطان خون^۵ جمع‌آوری شده است. این مجموعه داده شامل ۵۸ سلول منفرد و ۱۸ نوع جهش یکتا است. اطلاعات کامل در مورد این پایگاه داده از جمله، نام و نوع جهش‌های موجود در دیتابیس، نوع روش نمونه‌برداری و اطلاعاتی دیگر در پایگاه داده COSMIC در دسرس عموم قرار دارد. ماتریس ژنتوایپی این پایگاه داده شامل سه مقدار صفر، یک و دو می‌باشد که در آن صفر بیانگر عدم وجود جهش، یک بیانگر جهش هتروزیگوت و دو نمایانگر جهش هموزیگوت است. یکی از معایب این پایگاه داده نرخ بالای خطای توالی‌یابی تک‌سلولی و بالا بودن نرخ داده‌های از دست رفته (در حدود ۴۵ درصد کل داده‌ها) می‌باشد. همین امر سبب می‌شود تنوع درخت فیلوزنی نسبت داده شده به این پایگاه داده زیاد باشد. در واقع با در نظر گرفتن حالت‌های مختلف روابط دوبعدی جهش‌های گوناگون، می‌توان درخت‌های جهشی متنوعی از داده‌ها استنباط کرد.

۲.۲.۳ معیار ارزیابی

ارزیابی درختهای جهشی گوناگون از طریق روش LOOCV⁶. این روش همانند روش ارزیابی‌های متقابل^۷ با K قسمت می‌باشد با این تفاوت که در آن k برابر تعداد جهش‌ها (تعداد ستون‌های ماتریس ژنتوایپ) می‌باشد. در هر یک از درخت‌های استنباط شده، یکبار یک جهش حذف شده و میزان دقت مدل محاسبه می‌گردد. سپس این کار برای همه جهش‌های موجود تکرار می‌شود و در نهایت میانگین دقت مدل در حالت‌های مختلف محاسبه می‌شود و به عنوان دقت نهایی مدل گزارش می‌شود.

⁵Thrombocythemia

⁶Leave one out cross validation

⁷Cross validation

۳.۳ الگوریتم Bitphylogeny [۸۲]

این الگوریتم در سال ۲۰۱۵ ارائه شد و مانند الگوریتم کیم و سایمون از منطق بیزی بهره می‌برد. هدف این الگوریتم در کنار ساخت درخت جهشی تومور، پیدا کردن روابط بین کلون‌های مختلف درون یک تومور است. در داده‌های توالی‌یابی تکسلولی، بدلیل کمبود میزان نمونه‌گیری و در نتیجه محتمل بودن عدم حضور گونه‌های ژنومی جهش‌یافته در نمونه‌ها، برای تشخیص ناهمگنی‌های درون توموری باید رویکرد متفاوتی را برگزید. شاید یکی از دلایلی که هنوز از داده‌های توالی‌یابی آنبوه^۸ برای استنباط درخت فیلوزنی استفاده می‌شود همین باشد. در هر صورت در این مقاله سعی بر این است تا هر ۲ چالش زیر مورد بررسی قرار گیرد:

- تشخیص زیرنواحی یا کلون‌های درون یک تومور

- کشف روابط تکاملی کلون‌های درون یک تومور با یکدیگر

ماتریس ورودی (ماتریس ژنتوپی) این الگوریتم تعدادی سطر و ستون است که در آن سطراها بیانگر سلولها و ستونها نمایانگر انواع جهش‌های مختلف است. این ماتریس، یک ماتریس دودویی‌ها^۹ است که در آن بودن درایه z و زیانگر آن است که در سطر z ام جهشی از نوع z ام وجود ندارد. متعاقباً، اگر مقدار درایه z و z برابر یک باشد در سطر z ام جهشی از نوع z ام وجود دارد.

در این مقاله برای جستجو درختی که بیشترین تطابق با داده‌های ورودی را داشته باشد از الگوریتم زنجیره مارکوف مونت کارلو استفاده می‌شود. این الگوریتم سلول‌ها با ژنتوپ مشابه را درون یک گروه قرار می‌دهد و به این گروه‌ها کلون گفته می‌شود. در طی دسته‌بندی سلول‌ها کلون‌هایی ایجاد می‌شود که با احتمال زیاد توموری بوده ولی در نمونه‌گیری از بافت توموری حضور نداشته‌اند. شناسایی این گونه از کلون‌ها با توجه به روند گسترش و تکامل تومور، که به مرور زمان صورت می‌گیرد، امکان پذیر است. این الگوریتم قادر است تا یک تخمین زمانی از انتقال جهش از سطوح بالای درخت فیلوزنی به سطوح پایین‌تر را محاسبه کند. در این الگوریتم از داده‌های تغییرات تک نوکلئوتید استفاده شده‌است اما این روش این قابلیت را دارد تا بدون در نظر گفتن فرض مکان‌های بینهایت برای داده‌های متیلاسیون دی‌ان‌ای استفاده شود. از نکات قوت این الگوریتم می‌توان به محاسبه رخداد هر جهش در درخت فیلوزنی تومور اشاره کرد اما این مقدار احتمال بدلیل تعداد بالای داده‌های از دست رفته و

⁸Bulk sequencing

⁹Binary

نرخ بالای خطای مثبت کاذب و منفی کاذب^{۱۰}، بیش از مقدار واقعی است.

شایان ذکر است که این الگوریتم محدودیت‌های خاص خود را دارد. به عنوان مثال، در نظر گرفتن فرض مکان‌های بینهایت برای رخداد جهش‌ها و زمان محاسباتی بسیار بالا الگوریتم زنجیره مارکوف مونت کارلو برای استنباط درخت فیلوزنی از جمله این محدودیت‌ها می‌باشد. از دیگر محدودیت‌های این الگوریتم می‌توان به عدم تشخیص کلون‌های هموژنی و هتروژنی در یک نوع جهش از یکدیگر اشاره کرد. منظور از کلون‌های هموژنی در یک جهش معین آن است که اجزای تشکیل دهنده آن با توزیع یکنواخت در کنار یکدیگر قرار گرفته‌اند و این بدان معناست که احتمال رخداد هر جهش در این توده برابر با احتمال رخداد دیگر جهش‌هاست. در مقابل، یک توده دارای خاصیت هتروژنی است اگر اجزای تشکیل دهنده آن توزیع غیریکنواخت داشته باشد و به همین امر سبب می‌شود تا بدلیل حضور سلول‌های مختلف با توزیع گوناگون، احتمال رخداد جهش‌های مختلف متفاوت باشد.

۱.۳.۳ پایگاه داده

به منظور ارزیابی مدل استنباط کننده درخت تکاملی تومور، از دو پایگاه داده متفاوت در این مقاله استفاده شده است:

- دادگان مربوط به الگوهای متیلاسیون سرطان روده بزرگ
- دادگان شبیه‌سازی شده مربوط به سرطان خون

۲.۳.۳ معیار ارزیابی

به منظور ارزیابی عمکرد الگوریتم بیت‌فیلوزنی یک مقایسه بین خروجی این الگوریتم و خروجی‌های الگوریتم‌های خوشبندی k هسته‌ای^{۱۱} و دسته‌بندی سلسله مراتبی^{۱۲} صورت گرفته است. این مقایسه از طریق محاسبه معیار بیشینه عمق درخت تکاملی استنباط شده و مقدار درست‌نمایی صورت گرفته است. نتایج گزارش شده در این مقاله گواه از پایداری^{۱۳} و دقت^{۱۴} بسیار بهتر الگوریتم بیت‌فیلوزنی نسبت به دو الگوریتم دیگر است. در شکل

¹⁰False negative

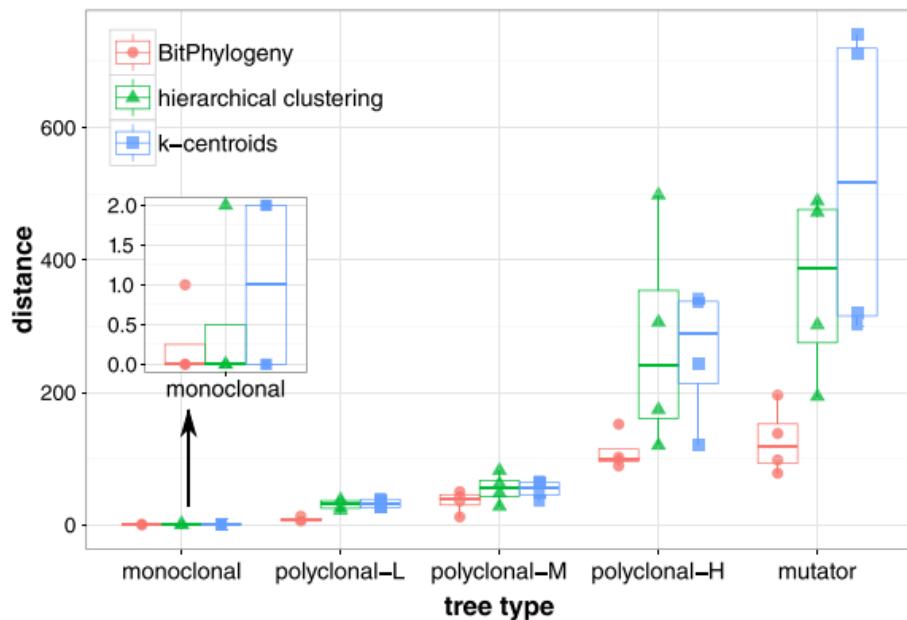
¹¹K-Centroids

¹²Hierarchical clustering

¹³Consistency

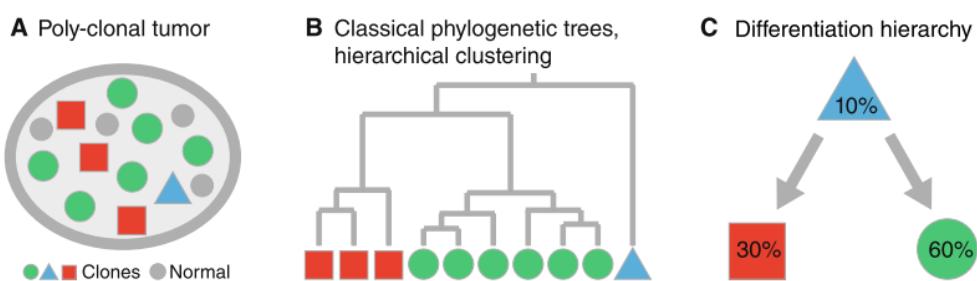
¹⁴Accuracy

۳.۲ میزان خطای عملکرد الگوریتم بیت‌فیلوجنی نسبت به دو الگوریتم خوشبندی k هسته‌ای و دسته‌بندی سلسله مراتبی در سطوح مختلف درخت در حالت‌های تک‌کلونی و چندکلونی قابل مشاهده است.



شکل ۳.۲: میزان خطای عملکرد الگوریتم بیت‌فیلوجنی نسبت به دو الگوریتم خوشبندی k هسته‌ای و دسته‌بندی سلسله مراتبی در سطوح مختلف درخت در حالت‌های تک‌کلونی و چندکلونی [۸۲]

در شکل ۴.۳ به طور کلی مراحل عملکرد الگوریتم بیت‌فیلوجنی را مشاهده می‌کنید. این شکل تومور چندکلونی A را نشان می‌دهد که به روش توالی‌یابی نمونه‌گیری شده است. این تومور شامل سه کلون مجزا و سلول‌های سالم (دایره‌های خاکستری رنگ) است. در تصویر میانی یک درخت بلقوه که نشان‌دهنده سیر تکاملی تومور است نشان داده شده است. در تصویر سمت راست درخت کلونی بدست آمده از درخت تکاملی تومور گفته شده با الگوریتم بیت‌فیلوجنی مشاهده می‌شود که در آن کلون‌ها و فراوانی هر یک مشهود است.



شکل ۴.۳: مراحل عملکرد الگوریتم بیت‌فیلوجنی

۴.۳ الگوریتم SCITE [۴۴]

این الگوریتم با استفاده از داده‌های توالی‌یابی تک‌سلولی^{۱۵} سعی در استنباط درخت فیلوزنی تومور دارد. همانطور که پیشتر نیز اشاره شد، یک تومور ناشی از تجمع تعدادی سلول با ویژگی‌های ژنی متفاوت است و این سلول‌ها سعی دارند تا این ویژگی‌های ژنی منحصر به فرد را از طریق تکثیر سلولی به سلول‌های بعدی منتقل کنند.

[۱۸]

وجود سلول‌ها با جهش‌های متفاوت سبب می‌شود که تومور از زیرنواحی گوناگون، که به کلون مشهور هستند، تشکیل شود. هر چه تومور از تعداد کمتری زیرکلون تشکیل شده باشد درمان آن ساده‌تر خواهد بود. در نظر گرفتن هر کلون به صورت یک تومور جداگانه، مطالعه و بررسی هر یک از این زیرتومورها به صورت دقیق‌تر و یافتن سیر تکاملی آنها سبب می‌شود تا درمان تومور به صورت کارآمدتری انجام شود. [۱۰]

یکی از چالش‌های بزرگ در زمینه تشخیص و مطالعه کلون‌های درون تومور، توالی‌یابی قسمت‌های مشترک دنباله‌های دی‌ان‌ای است، زیرا شامل ترکیب‌های بسیار زیادی (در حدود میلیون‌ها ترکیب) از ژنهای سلول‌های گوناگون است. جهش‌های بدست آمده از ترکیب توالی سلول‌های مختلف، با تعداد زیرنواحی توموری (کلون) متناسب است و با استفاده از تعداد زیرنواحی می‌توان تخمین نزدیکی از جهش‌های درون یک نمونه را بدست آورد [۵۳]. به همین دلیل به منظور شناسایی دقیق هر یک از زیرنواحی توموری (کلون‌ها) لازم است تا اطلاعات حاصل از نواحی مشترک کلون‌ها به دقت مورد تحلیل و تجزیه قرار گیرد. [۶۲]

الگوریتم Scite از طریق داده‌های توالی‌یابی تک‌سلولی قادر است سیر تکاملی تومور را از طریق درخت جهشی تومور که در آن ترتیب وقوع جهش‌ها مشخص است یا از طریق استنباط درخت فیلوزنیک تومور که در آن هر برگ نشان دهنده یک سلول است، نشان دهد. خروجی مدل Scite نتیجه ارزیابی بهتری در مقایسه با الگوریتم بیت‌فیلوزنی بر روی داده‌های واقعی داراست. الگوریتم Scite از طریق معیار بیشینه درست‌نمایی و احتمال رخداد هر جهش و با استفاده از ماتریس ژنتایپ ورودی تعیین می‌کند که کدام درخت استنباط بهتری از سیر تکاملی تومور است. در حالتی که تعداد جهش‌ها بسیار زیاد باشد یعنی تعداد ستون‌های ماتریس ژنتایپ ورودی زیاد باشد، ساخت درخت فیلوزنیک راحت‌تر خواهد بود، اما در حالتی که تعداد سلول‌ها زیاد باشد (تعداد سطرهای ماتریس ژنتایپ بالا باشد) ساخت درخت جهشی تومور (ترتیب وقوع جهش‌ها) راحت‌تر است. به طور خلاصه اینکه کدام نوع درخت (جهشی یا فیلوزنیک) در نهایت بیان‌کننده سیر تکاملی تومور باشد به نز

¹⁵Single cell sequencing

جهش‌های توموری و روش توالی‌یابی داده‌ها بستگی دارد.
در الگوریتم Scite از دو فرض اصلی استفاده می‌شود:

- فرض مکان‌های بی‌نهایت^{۱۶} که بر طبق آن هر جهش تنها یکبار در هر موقعیت از ژنوم رخ می‌دهد.
- فرض جهش‌های نقطه‌ای یعنی مدل تکاملی تومور به جهش‌های نقطه‌ای محدود می‌شود.

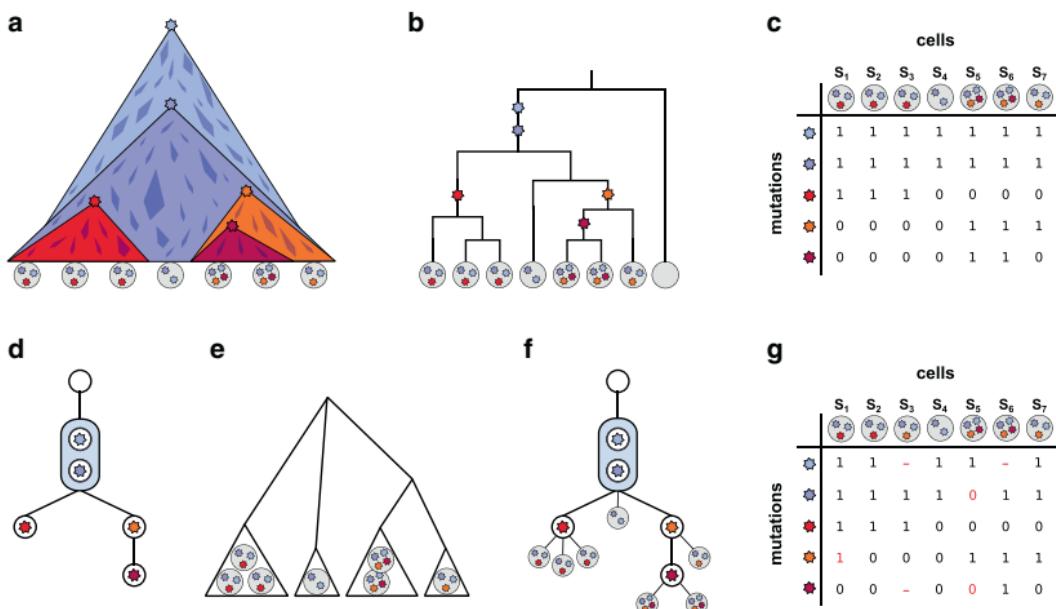
مانند الگوریتم بیت‌فیلوزنی از یک ماتریس ژنوتایپ (ماتریس دودویی‌ها که سطرها نمایانگر نمونه‌ها و ستون‌ها بیانگر جهش‌های است) به عنوان ورودی الگوریتم استفاده می‌شود. موقعیت هر جهش به صورت درایه π و ز از ماتریس $m \times n$ مشخص می‌شود. به این صورت که مقدار صفر در سطر π ام و درایه π زام بیانگر آن است که جهش از نوع π در سطر π وجود ندارد. یک ماتریس ژنوتایپ را ماتریس فیلوزنی کامل^{۱۷} گوییم هر گاه به ازای آن یک درخت فیلوزنیک متناظر باشد. در الگوریتم scite همانند الگوریتم بیت‌فیلوزنی از الگوریتم زنجیره مارکوف مونت کارلو برای اسنباط درخت تکاملی تومور از داده‌های توالی‌یابی تک سلولی استفاده می‌شود، با این تفاوت که فضای جستجو برای انتخاب پارامترها بسیار محدودتر از حالت بیت‌فیلوزنی است و نرخ خطاهای داده (مثبت کاذب و منفی کاذب) برای همه جهش‌های یکسان در نظر گرفته شده است. محدود کردن فضای جستجو برای انتخاب پارامترها از طریق نمونه‌برداری در این فضای سبب می‌شود تا بر اساس سیر زمانی جهش، بیشینه درست‌نمایی از روی توزیع احتمال پیشین نمونه‌ها بدست آید. یکی از مزایای این روش محاسبه نرخ خطای توالی‌یابی است. شکل ۵.۳ یک استنتاج تکاملی از داده‌های توالی‌یابی تک سلولی را نشان می‌دهد. در این شکل، a یک ماتریس فیلوزنی کامل است اما g ماتریس داده‌های واقعی است که شامل مقادیر از دست‌رفته، خطای مثبت کاذب و خطای منفی کاذب است. ماتریس داده‌های واقعی با D و ماتریس فیلوزنی کامل را با E نشان داده شده است.

منظور از خطای مثبت کاذب این است که به عنوان مثال در یک موقعیت خاص از ماتریس E جهشی وجود ندارد (مقدار ماتریس برابر صفر است) اما در همین موقعیت مقدار یک (وجود جهش) در ماتریس D وجود دارد. نرخ خطای مثبت کاذب با α و نرخ خطای منفی کاذب با β نشان داده می‌شود. مقادیر α و β از طریق روابط ۴.۳ تعریف می‌گردند.

$$\begin{aligned} P(D_{ij} = 1 | E_{ij} = 0) &= \alpha, & P(D_{ij} = 0 | E_{ij} = 0) &= 1 - \alpha \\ P(D_{ij} = 0 | E_{ij} = 1) &= \beta, & P(D_{ij} = 1 | E_{ij} = 1) &= 1 - \beta \end{aligned} \quad (4.3)$$

¹⁶Infinite sites

¹⁷Perfect phylogenetic matrix



شکل ۵.۳: یک استنتاج تکاملی از داده‌های توالی‌بایی تک‌سلولی [۴۴]

در این معادلات فرض بر استقلال نرخ خطاهای مشاهده شده است. مقدار درست‌نمایی درخت جهشی T با بردار ضمیمه θ و نرخ خطای $(\alpha, \beta) = \theta$ به صورت زیر محاسبه می‌گردد.

$$y = xxxxxxxxxxxxxxxxxxxxxxxxx \quad (5.3)$$

در معادله بالا E ماتریس جهش‌دار است که با درخت جهشی T و بردار ضمیمه θ تعریف می‌گردد. توزیع احتمال پسین به صورت زیر محاسبه می‌گردد:

$$y = xxxxxxxxxxxxxxxxxxxxxxxxx \quad (6.3)$$

به منظور بالا رفتن سرعت همگرایی مدل زنجیره مارکوف مونت کارلو فرض می‌شود که بردار ضمیمه θ توزیع یکنواخت دارد. در نتیجه:

$$y = xxxxxxxxxxxxxxxxxxxxxxxxx \quad (7.3)$$

اندازه فضای جستجو برای دو پارامتر درخت جهشی T و بردار ضمیمه ∂ برابر با $(n+1)^{m-1} \times (n+1)^m$ می‌باشد. این فضای جستجو با فرض یکنواخت بودن توزیع بردار ضمیمه ∂ و طبق معادله بالا و حذف بردار ضمیمه به $(n+1)^{m-1}$ انتخاب کاهش می‌یابد. پس از همگرایی با استفاده از الگوریتم زنجیره مارکوف مونت کارلو و احتمال پسین، بهترین ترکیب درخت جهشی T با بردار ضمیمه ∂ با بیشینه درست‌نمایی بدست می‌آید:

$$y = xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx \quad (8.3)$$

منظور از MAP در این معادله حالتی است که بیشینه درست‌نمایی رخ داده است.

۱.۴.۳ پایگاه داده

به منظور ارزیابی عملکرد الگوریتم Scite برای استباط درخت تکاملی تومور از داده‌های توالی‌یابی تک سلولی از داده‌های واقعی و شبیه‌سازی شده، استفاده شده است. مجموعه داده‌های استفاده شده جهت ارزیابی الگوریتم عبارتند از:

- داده‌های توالی‌یابی تک سلولی از یک نمونه تومور مغز استخوان با ۵۸ سلول سرطانی و ۱۸ نوع جهش با نرخ خطای مثبت کاذب 4×10^{-4} و نرخ خطای منفی کاذب 9×10^{-4} .
- داده‌های توالی‌یابی تک سلولی یک نوع خاص از سرطان کبد با ۱۷ سلول سرطانی و ۵۰ نوع جهش با مقادیر نرخ خطای مثبت کاذب 5×10^{-5} و نرخ خطای منفی کاذب 3×10^{-5} و نرخ داده‌های از دست رفته ۲۲ درصد.
- داده‌های توالی‌یابی تک سلولی نمونه‌گیری شده از سرطان سینه با ۴۷ سلول سرطانی و ۴۰ نوع جهش و با نرخ خطای ترک آلل 3×10^{-6} درصد و نرخ خطای مثبت کاذب 6×10^{-6} .

شایان ذکر است که مدت زمان استباط یک درخت فیلورژنی تا حد زیادی به پیچیدگی داده‌های ورودی بستگی دارد بطوریکه برای ساخت یک درخت با ۵۰ تا ۱۰۰ سلول، مدت زمانی در حدود چندین دقیقه طول می‌کشد. از مهمترین محدودیت‌های این الگوریتم می‌توان به فرض مکان‌های بی‌نهایت اشاره کرد، زیرا این امکان وجود دارد که در یک محل مشخص از یک دنباله دی‌ان‌ای، یک جهش مشخص چندین بار رخداد و یا در محل‌های مختلف

از یک دنباله ژنی جهش‌های مشابه رخ دهد که این موارد در فرض مکان‌های بی‌نهایت در نظر گرفته نمی‌شود. از دیگر محدودیت‌های این روش آن است که جهش‌هایی که در همه سلول‌ها وجود دارند یا جهش‌هایی که فقط در یک سلول مشاهده شده‌اند (سطری با مقادیر تماماً یک در ماتریس ورودی) در روند استنباط درخت مورد استفاده قرار نمی‌گیرند.

۵.۳ الگوریتم [۵۸] Onconem

این الگوریتم در سال ۲۰۱۶ با هدف یافتن تاریخچه تکاملی ناحیه‌های درون توموری با استفاده از داده‌های توالی‌یابی تک‌سلولی ارائه گردید. این الگوریتم قادر است تا ناحیه‌های درون توموری مشابه را درون یک دسته قرار دهد و برای آنها یک ژنتوتایپ یکتا در نظر بگیرد. این الگوریتم بر مبنای تغییرات تک نوکلئوتیدی، درخت تکاملی تومور را استنباط می‌کند و قادر به یافتن خطاهای ژنتوتایپی می‌باشد. در نهایت با ارزیابی بر روی داده‌های آزمایش، مدل نهایی سنجیده شده و سلول‌ها با جهش‌های یکسان در یک گروه دسته‌بندی شده و در انتهای رابطه میان جهش‌ها و ژنتوتایپ‌های مشاهده شده و مشاهده نشده (پیش‌بینی شده) مشخص می‌گردد. این الگوریتم هم می‌تواند درخت کلونال توموری و هم درخت فیلوزنیک توموری (قرارگرفتن سلول‌ها به عنوان برگهای درخت) را به عنوان خروجی بدست دهد. ورودی این الگوریتم ماتریس دودوبی ژنتوتایپ به همراه نرخ خطاب مثبت کاذب و نرخ خطاب منفی کاذب و نرخ خطاب داده‌های از دست رفته است. در ادامه، الگوریتم سعی می‌کند تا سلول‌ها با ژنتوتایپ‌های مشابه را در یک گروه قرار دهد و در نهایت درختی که بیشترین شباهت را با دسته‌بندی صورت گرفته را دارد به عنوان درخت تکاملی تومور استنباط کند. از نکات قوت این الگوریتم آن است که قادر است کلون‌هایی را که احتمال وجود آنها بالاست اما در داده‌های نمونه‌گیری شده حضور ندارند حدس بزند. این الگوریتم از دو قسمت اصلی تشکیل شده است:

- ایجاد یک مدل احتمالاتی به منظور مدل کردن جمعیت جهش‌ها بر مبنای داده‌های نویزی و روابط میان

داده‌ها

- پیدا کردن درخت‌هایی با بیشترین میزان درست‌نمایی در فضای جستجو

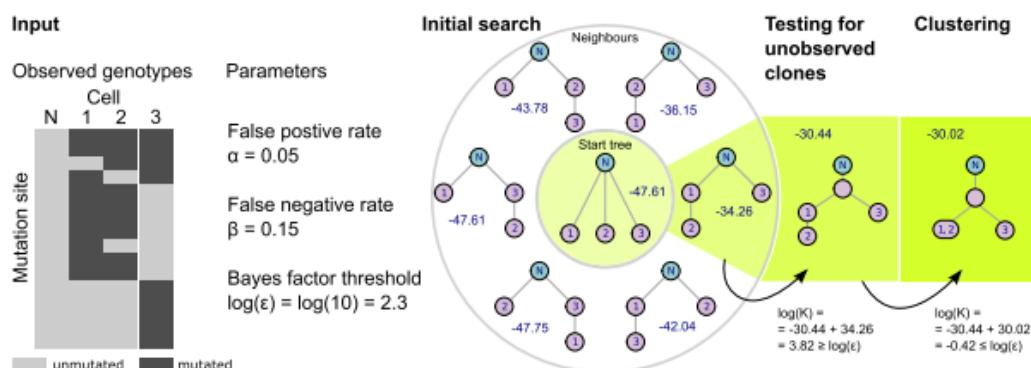
توزیع احتمال پسین با فرض D به عنوان مجموعه داده‌ای مدل به صورت زیر محاسبه می‌گردد

$$y = xxxxxxxxxxxxxxxxxxxxxxxx \quad (9.3)$$

که در آن τ نمایانگر یک درخت جهش‌دار (که نباید حتماً دودویی باشد) است که ریشه آن یک گره سالم و بدون جهش است و θ یک بردار رخداد است. در این رابطه فرض بر آن است که $(\tau) p$ دارای توزیع یکنواخت است. رابطه بالا می‌تواند به شکل زیر بازنویسی شود

$$y = xxxxxxxxxxxxxxxxxxxxxxxx \quad (10.3)$$

بر طبق این رابطه، برای درختی با n راس، فضای جستجو شامل n^{n-2} انتخاب است که هزینه محاسباتی بسیار بالایی برای درختانی با راس‌های بیشتر از ۹ دارد. طبق شکل ۹.۳، برای محدود کردن فضای جستجو از یک الگوریتم اکتشافی استفاده می‌شود تا اطمینان حاصل شود که خروجی الگوریتم یک نقطه بینه‌محمل نباشد. نقطه قوت این الگوریتم سرعت بالای استنباط درخت برای داده‌های کم است ولی در مقابل از محدودیت‌های آن می‌توان به فرض بینهایت اشاره کرد.



شکل ۶.۳: نمای شماتیکی از الگوریتم [۵۸] Onconem

رونده کلی الگوریتم Onconem در شکل بالا توضیح داده است. طبق این شکل، ماتریس دودویی ژنتوپاپی به همراه نرخ خطاهای α و β به عنوان ورودی الگوریتم استفاده می‌شوند. طبق شکل بالا میزان درست‌نمایی اولیه برابر $61/47$ – محاسبه شده است اما از میان همه درخت‌های همسایه درخت اولیه، آن درختی که بیشترین

درست‌نمایی را دارد به عنوان درخت، اولیه انتخاب می‌شود (با درست‌نمایی ۲۶، ۳۶). در ادامه یک گرهای که احتمال رخداد آن طبق ماتریس ورودی بالاست ولی در داده‌های ورودی وجود ندارد به درخت اضافه می‌شود. در این حالت مقدار درست‌نمایی به $3/82$ افزایش می‌یابد و این کلون مشاهده نشده بدلیل بزرگتر بودن مقدار درست‌نمایی از آستانه تعیین شده، به مدل افزوده می‌شود. در نهایت گرهای یک شاخه تا جایی که سبب کاهش میزان درست‌نمایی نشوند، در یک کلون تجمعی می‌شوند.

۱.۵.۳ پایگاه داده

به منظور ارزیابی عملکرد الگوریتم Onconem از دو پایگاه داده مجزا استفاده شده است

- داده‌های توالی یابی تک سلولی مربوط به سرطان مثانه که شامل ۴۴ سلول سرطانی است. در حدود ۵۵ درصد از انواع جهش‌های موجود در این پایگاه داده، اطلاعاتی در دسترس نیست یعنی بیش از نیمی از داده‌های موجود از اطلاعات از دست رفته^{۱۸} اند. خروجی الگوریتم Onconem برای این پایگاه داده یک درخت فیلوزنی با سه کلون اصلی می‌باشد و یک چهارم سلول‌های جهش‌یافته را شامل می‌شود.
- داده‌های مربوط به سرطان خون که در مدل کیم و سایمون و الگوریتم بیتفیلوزنی از آن استفاده شده بود، در این ارزیابی مورد استفاده قرار گرفت. میزان لگاریتم درست‌نمایی الگوریتم Onconem برای این مجموع داده برابر ۹۹۶۴ – گزارش شده است که بالاتر از مقداری است که الگوریتم بیتفیلوزنی به آن رسیده بود (۱۱۵۸۴).

۶.۳ الگوریتم [۵۸] Sasc

سرطان ناشی از جهش‌های ژنومیک یک سلول است که این جهش‌ها به مرور زمان رشد و تکثیر می‌یابند و زیرنواحی متفاوتی را ایجاد می‌کنند. این زیرنواحی، که به آنها کلون نیز گفته می‌شود، خصوصیات متفاوتی دارند و در کنار هم یک توده سرطانی را تشکیل می‌دهند. بررسی تاریخچه تکاملی تومور می‌تواند کارآمدی درمان‌های موجود را بهبود بخشد و امکان عود مجدد تومور را تا حد زیادی کاهش دهد. به منظور درک بهتر تاریخچه تکاملی

¹⁸Missing data

تومور فرض‌های گوناگونی جهت ساده‌سازی مسئله صورت می‌گیرد، مثل فرض مکان‌های بی‌نهایت که طبق آن هر جهش یکتاپی تنها یکبار رخ می‌دهد. مطالعات زیادی صورت گرفته است که نشان می‌دهد در نظر گرفتن فرض مکان‌های بی‌نهایت به تنها ی برای استنباط روند تکاملی تومور کافی نیست و محدودیت‌هایی دارد، به همین منظور برای درک بهتر نواحی ناهمگن توموری باید فرض‌های دیگری را به مسئله اضافه کنیم. به همین دلیل یک فرضیه جدید تحت عنوان k -dollo ارائه گردید که بر طبق آن و بر خلاف فرض مکان‌های بی‌نهایت، هر جهشی تنها یکبار رخ می‌دهد اما امکان از دست دادن این جهش به تعداد k در تاریخچه تکاملی تومور وجود دارد. الگوریتم Sasc که در سال ۲۰۱۸ ارائه گردید، از اولین الگوریتم‌هایی بود که از فرض k -dollo جهت استنباط درخت تکاملی تومور بهره برد. به مانند الگوریتم Onconem، این الگوریتم به منظور محدود کردن فضای جستجو از یک الگوریتم درخت اکتشافی بهره می‌برد. الگوریتم اکتشافی استفاده شده در این روش، الگوریتم شبیه‌سازی ذوب فلزات است و هدف آن پیدا کردن بیشینه درست‌نمایی برای تابع احتمال رخداد پسین در فضا جستجو است. طبق این الگوریتم، ابتدا از طریق مجموعه‌ای از انتخاب‌های نمونه‌برداری شده از فضا جستجو یک راه حل برای مسئله ارائه می‌گردد. اگر مقدار درست‌نمایی نسبت به حالت اولیه بهبود یافته بود، با احتمال یک پذیرفته می‌شود در غیر این صورت احتمال رخداد آن حالت صفر در نظر گرفته می‌شود. این الگوریتم سعی دارد تا بیشینه درست‌نمایی ماتریس ژنوتایپ ورودی را حساب کند. ورودی این الگوریتم در کنار ماتریس ژنوتایپ، نرخ خطای مشتبه کاذب، نرخ خطای منفی کاذب و نرخ خطای اطلاعات از دست رفته است و بیشینه درست‌نمایی از رابطه زیر بدست می‌آید:

$$y = xxxxxxxxxxxxxxxxxxxxxxxxx \quad (11.3)$$

۱.۶.۳ پایگاه داده:

به منظور ارزیابی عملکرد الگوریتم Sasc از دو پایگاه داده معجزا استفاده شده است:

- داده‌های توالی‌یابی تک سلولی سرطان مثانه

- داده‌های شبیه‌سازی شده سرطان خون

خروجی الگوریتم در مقایسه با الگوریتم Scite از مقدار بیشینه درست‌نمایی بیشتری برای مدل کردن دادها برخوردار است.

۷.۳ الگوریتم Scarlet [۶۴]

مدل ارائه شده در این مقاله که در سال ۲۰۲۰ به چاپ رسیده است، یک مدل تکاملی است که امکان حذف هر نوع جهشی را با در نظر گرفتن حذف خطای^{۱۹} در نظر می‌گیرد. این مدل اجازه حذف دگرگونی تک‌هسته‌ای^{۲۰} را تنها هنگامی که با شواهد داده توالی‌یابی تک‌سلولی دی‌ان‌ای از یک حذف در همان مکان هندسی^{۲۱} همراه شده باشد، می‌دهد. این مدل پایه الگوریتم اسکارلت خواهد بود که فیلوزنی تومور را از داده توالی‌یابی تک‌سلولی دی‌ان‌ای با احتساب هر دوی خطای توالی‌یابی و حذف جهش‌ها نتیجه می‌دهد. تعداد کمی از جهش‌های سوماتیک منجر به پیشروی سرطان می‌شوند، اما تمام جهش‌های سوماتیک نشانگرهای زیستی تاریخچه تکامل تومور هستند. روش‌های غالب ساخت فیلوزنی داده توالی‌یابی تک‌سلولی دی‌ان‌ای از دگرگونی تک‌هسته‌ای‌ها به عنوان نشانگرهای زیستی استفاده می‌کنند اما در به حساب آوردن تغییر تعداد کپی، که ممکن است با دگرگونی تک‌هسته‌ای همپوشانی داشته باشد و منجر به حذف دگرگونی تک‌هسته‌ای شود، ناتوان است. الگوریتم پیشنهادی اسکارلت، فیلوزنی تومور را از داده توالی‌یابی تک‌سلولی دی‌ان‌ای، خطای توالی‌یابی و حذف دگرگونی تک‌هسته‌ای از طریق تغییر تعداد کپی را لحاظ می‌کند. این الگوریتم عملکرد بهتری نسبت به روش‌های موجود بر روی داده‌های شبیه‌سازی شده دارد. توالی‌یابی تک‌سلولی دی‌ان‌ای از تومور بدلیل افزایش بازدهی الگوریتم و کاهش هزینه ایزوله کردن، نشانه‌گذاری و توالی‌یابی سلول‌های انفرادی از محبوبیت روزافزونی برخوردار است.

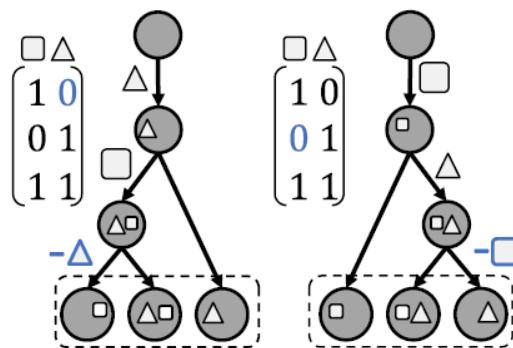
شکل ۷.۳ فیلوزنی با فرض دولو^{۲۲} را نشان می‌دهد. این مدل با شناسایی حذف جهش‌ها به منظور رفع تناقض مدل مکان‌های بی‌نهایت می‌تواند درخت‌های ۷.۳ را بسازد. هر دو مدل دولو و مکان‌های بی‌نهایت می‌توانند چندین درخت ممکن را بسازند. حتی در حالت‌های ساده‌ای که خطای وجود ندارد، استنباط چندین فیلوزنی سازگار با داده‌ها ممکن است وجود داشته باشند. در صورتی که خطای وجود داشته باشد و عدم قطعیت در ماتریس جهش وجود داشته باشد، تعداد این درخت‌های احتمالی بسیار بیشتر خواهد شد. خطای داده توالی‌یابی

¹⁹loss-supported

²⁰Single nucleotide variant (SNV)

²¹Loci

²²Dollo

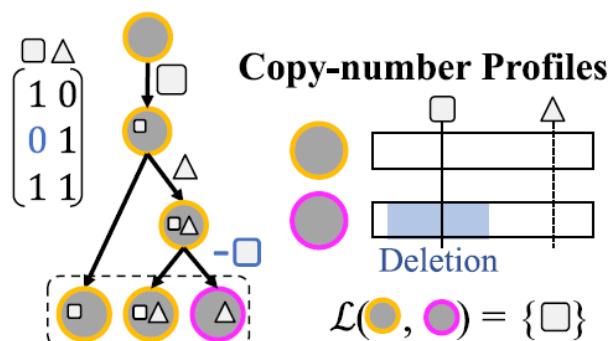


شکل ۷.۳: عنواننتنتنتنتنتنتنتنتنتنت

تک سلولی دی‌ان‌ای و حذف جهش‌ها منجر به پیچیدگی مسئله و ابهام در استنباط فیلوزنی خواهد شد. به عنوان مثال با مشاهده کردن ^{۲۳} در ماتریس جهش به جای ۱ نمی‌توان براحتی بین خطاهای داده‌ها و حذف جهش‌ها تقاضی قائل شد. عمدۀ محدودیت الگوریتم‌های دولو یا مکان‌های بی‌نهایت‌ایی که اجازه حذف جهش‌ها را می‌دهند این است که هیچ‌کدام از این روش‌ها شواهد تغییر تعداد کپی در حذف جهش‌ها را در یک مکان هندسی در نظر نمی‌گیرند. مدل‌های چند حالته ^{۲۴} از تکامل تومور که از داده‌های توالی‌یافته با نمونه‌های زیادی از تومور استفاده می‌کنند. این نگرش‌ها نه خطای موجود در داده توالی‌یابی تک سلولی دی‌ان‌ای را مدل می‌کنند و نه در ابعاد صدھا یا هزاران سلول قابلیت مدل کردن را دارند.

از آنجایی که حذف جهش‌ها پیچیده‌ترین قسمت در تکامل دگرگونی تک‌هسته‌ای است و مسئول اکثر تناقضات در مدل مکان‌های بی‌نهایت در داده‌های توالی‌یابی تک سلولی دی‌ان‌ای هستند، در نگرش ارائه شده در این الگوریتم، حذف جهش‌ها را با استفاده از داده‌های جهش‌های تغییر تعداد کپی از همان سلول‌ها محدود خواهد کرد. در نتیجه الگوریتم اسکارلت با یکپارچه کردن دگرگونی تک‌هسته‌ای و داده‌های حذف و تغییر تعداد کپی ^{۲۵}، درخت فیلوزنی را براساس داده توالی‌یابی تک سلولی دی‌ان‌ای می‌سازد. الگوریتم اسکارلت براساس مدل فیلوزنی با در نظر گرفتن حذف خطای است که حذف جهش‌ها را محدود به مکان‌های هندسی خواهد کرد. به عنوان مثال در این الگوریتم داده تغییر تعداد کپی گواه یک حذف است. شکل زیر مدل فیلوزنی با در نظر گرفتن خطای حذف را نشان می‌دهد که با استفاده از داده تغییر تعداد کپی سعی در محدود کردن حذف جهش‌ها دارد تا بتواند ابهام ^{۲۶} ایجاد شده را رفع کند.

²³Multi-state²⁴Copy number variation (CNV)²⁵conflict



مدل فیلوژنی با در نظر گرفتن حذف خطای مدلی از تکامل دگرگونی تک هسته‌ای است که جهش حداکثر یکبار رخ خواهد داد (۰-۵۱) اما حذف جهش‌ها (۱-۵۰) توسط مجموعه از مقدار خطای حذف که توسط تغییر تعداد کپی‌ها تعریف می‌شوند محدود خواهد شد. برای هر جفت سلول، از مجموعه جهش‌های تغییرات تعداد کپی، مجموعه خطای به صورت ، تعریف خواهد شد. مدل فیلوژنی با در نظر گرفتن حذف خطای توسعه دهنده مدل‌های مکان‌های بی‌نهایت و دلولو می‌باشد. ضمناً الگوریتم اسکارلت متکی بر مدل احتمالاتی تعداد خوانش‌ها برای هر دگرگونی تک هسته‌ای است تا خطاهای و داده‌های از بین رفته، که در توالی یابی تک سلولی دی‌ان‌ای معمول هستند، را مورد توجه قرار می‌دهد.

اسکارلت سہ ویژگی، مہم دارد:

- مدل فیلوژنی با در نظر گرفتن حذف خطای حذف جهش ها را محدود به مکان هایی می کند که کاهش متناظر با آن در تعداد جهش های کپی وجود داشته باشد.
 - این الگوریتم با استفاده از مدل فیلوژنی با در نظر گرفتن حذف خطای ابتدا درخت فیلوژنی اولیه استنتاج شده را پایش و سپس از طریق داده های تغییر تعداد کپی، فیلوژنی نهایی را استنباط می کند.
 - استنتاج مبتنی بر بیشینه درست نمایی از دگرگونی های تک هسته ای با استفاده از مدل احتمالاتی تعداد خوانش های مشاهده شده در داده های توالی یابی تک سلولی دی ان ای توسط الگوریتم اسکارلت اجرا می شود.

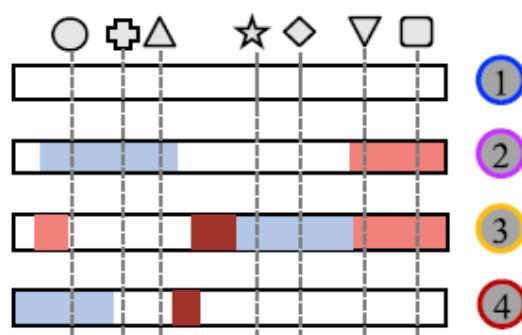
اگر تعدادی حهش‌های حذف و تغییر تعداد کیه، موجود باشند و وضعیت حهش‌های تغییر تعداد کیه، را با

رنگ قرمز و حذف نواحی ژنوم در طول کل ژنوم را با رنگ آبی نشان دهیم:

CNAs

Copy-number profiles

Mutations



شكل ٩.٣: عناوين ترتيبية

آنگاه مجموعه خطای مدل فیلوزنی با در نظر گرفتن حذف خطای توسط مجموعه های ۱۱.۳ نمایش داده خواهد

شل:

Supported losses \mathcal{L}

$$\mathcal{L}(1,2) = \{\textcircled{O}, \textcolor{red}{\textbullet}, \triangle\}$$

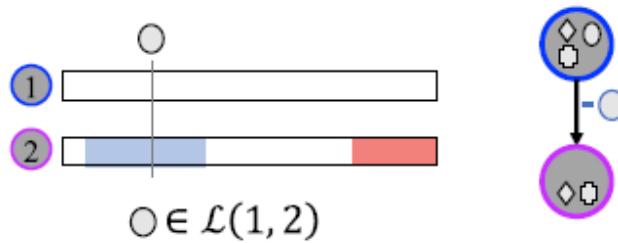
$$\mathcal{L}(1,4) = \{\textcircled{O}\}$$

$$\mathcal{L}(2,3) = \{\star,\diamond\}$$

شکا ۱۰.۳: عنواننامه

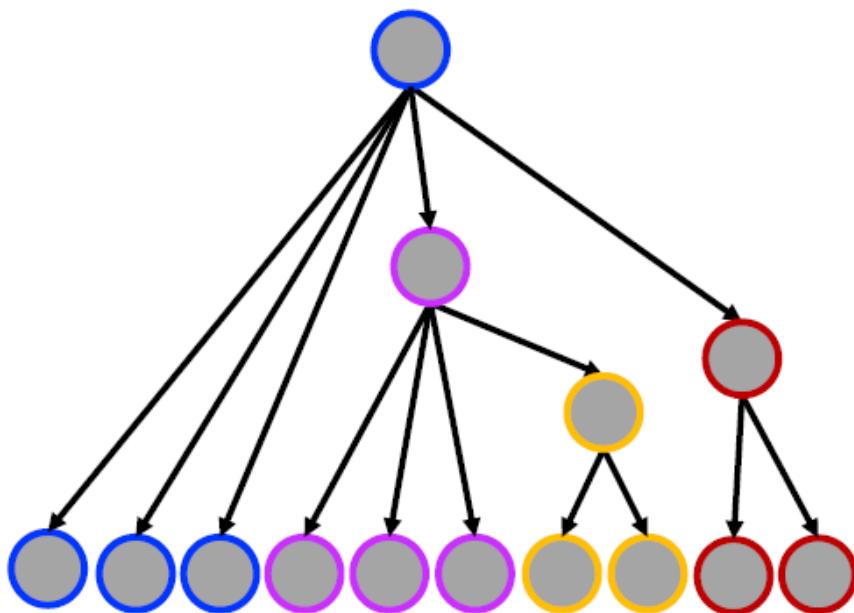
الگوریتم اسکارلت به صورت مستقیم وضعیت جهش‌های حذف و تغییر تعداد کپی سلول‌های پدری را نشان نخواهد داد. به منظور غلبه بر این موضوع یک درخت برای جهش‌های حذف و تغییر تعداد کپی زیر را در نظر نماید.

۹۳-۱۹۰۶-۲۵، نای، الگوریتم اسکارلت در نظر گرفته و شده.



شکل ۱۱.۳: عنواننتننتننتننتن

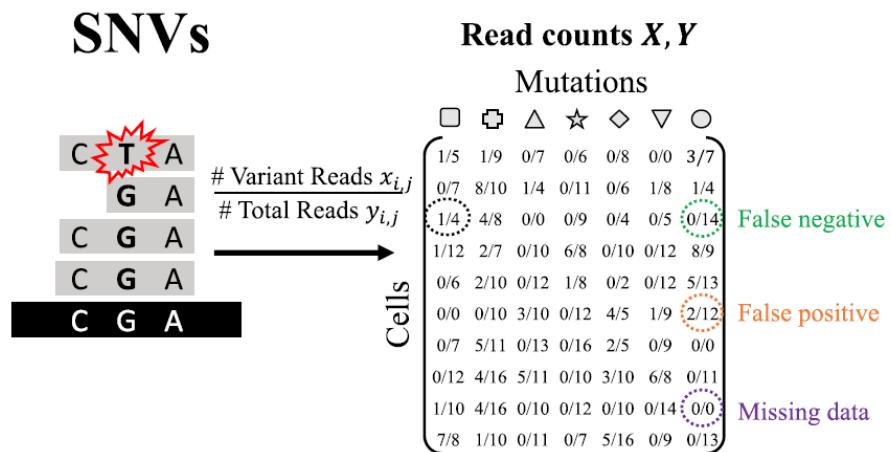
Copy-number tree T



شکل ۱۲.۳: عنواننتننتننتننتن

- مجموعه خطاهای ناشی از حذف جهش‌ها است، که مجموعه‌های تهی در آن نمایش داده نمی‌شوند. این مجموعه جهش‌هایی که تحت تاثیر حذف قرار می‌گیرند را نشان می‌دهند.
- یک درخت فیلوزنی برای جهش‌های تغییر تعداد کپی، که با استفاده از آن می‌توان روابط بین سلول‌های مشاهده شده (برگ‌ها) را آنگونه که توسط وضعیت جهش‌های تغییر تعداد کپی تعیین شده، نشان داد.

برای دگرگونی‌های تک‌هسته‌ای تنو^{۲۶} X و مجموع^{۲۷} Y از تعداد خوانش‌ها^{۲۸} برای هر سلول و هر جهش مطابق ماتریس^{۱۳.۳} تهیه شده است:



شكل ١٣.٣: عنوان

در ادامه الگوریتم اسکارلت، روابط بین اتصال سلولها (T) را از سلول های مشاهده شده(برگها) و ماتریس جهش بیشینه درست نمایی^{*} B را با محدود کردن حذف جهش ها به مجموعه

از خطاهای احتمالی حساب میکند. سپس با مقایسه T از $b_{i,j}$ و انتخاب بیشینه درستنمایی B^* را با استفاده از مدل احتمالاتی برای حضور ($1 = b_{i,j}$) و یا عدم حضور ($0 = b_{i,j}$) هر دگرگونی تک هسته‌ای در هر سلول را انجام می‌دهد.

مقایسه T از T'

مدل احتمالاتی برای توالی پابجی داده:

ساختن درخت اتصالات T:

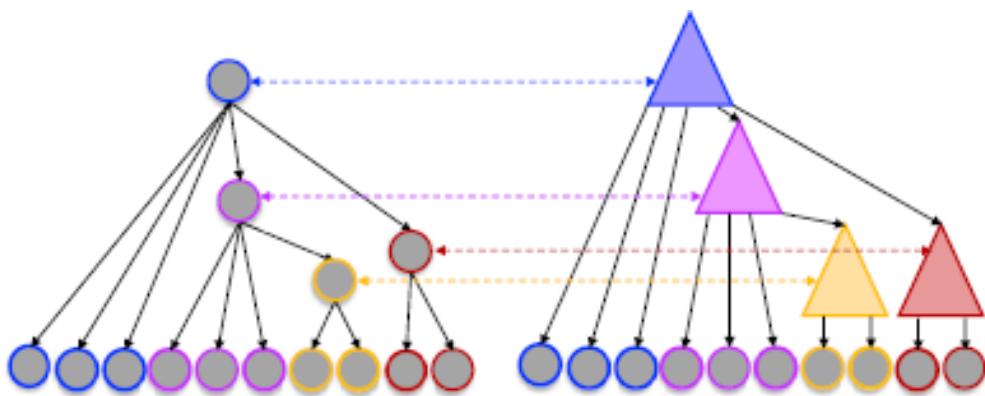
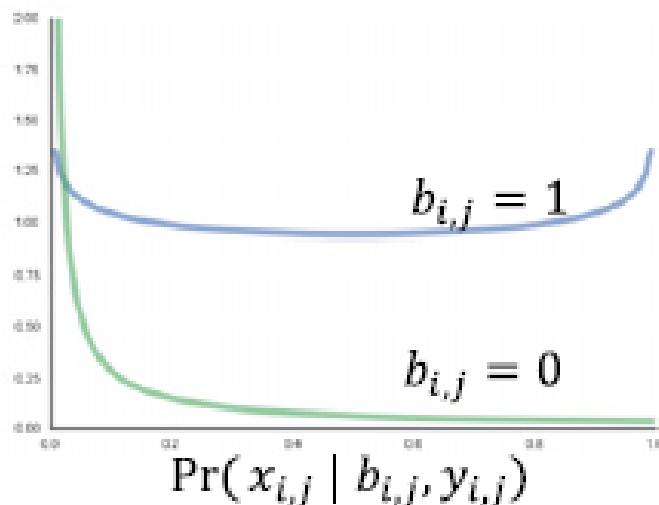
و در نهایت ماتریس جهش‌ها B^* با پیشینه درست‌نمایی:

الگوریتم اسکارلت باید مسئله بیشینه درست‌نمایی همراه با انتخاب بهترین حذف‌هارا حل کند. این الگوریتم از طریق یافتن ماتریس جهش با بیشینه درست‌نمایی B^* انجام خواهد گرفت. در اینجا $L(T)$ مجموعه برگ‌های

26 Variant

27 Total

Total ²⁸Read counts

شکل ۱۴.۳: مقایسه T از T' 

شکل ۱۵.۳: مدل احتمالاتی برای توالی یابی داده

درخت T را بیان می‌کند.

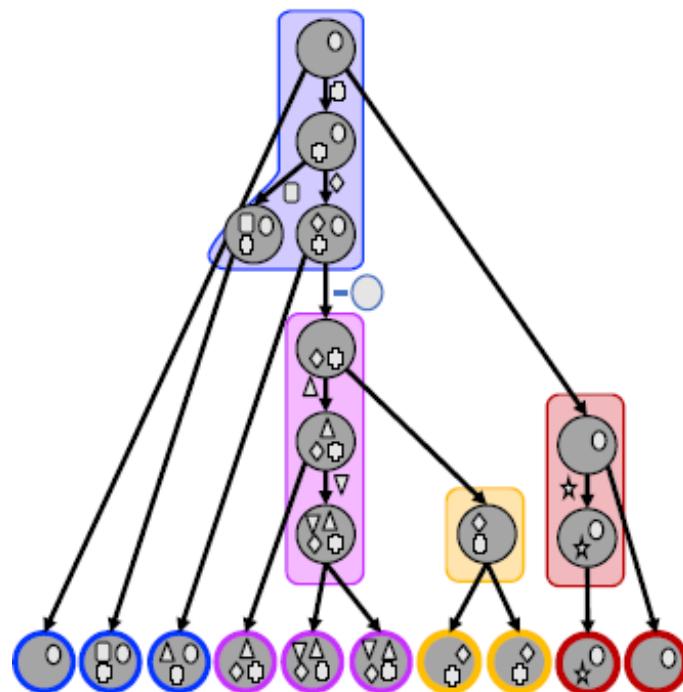
$$y = xxxxxxxxxxxxxxxxxxxxxxx$$

الگوریتم اسکارلت از ۲ قسمت اصلی زیر تشکیل شده است.

- محاسبه وضعیت‌های جهش با بیشینه درست‌نمایی R^* از ریشه زیردرخت‌ها^{۲۹}

- استنتاج هر زیردرخت به صورت مستقل با هدف بیشینه درست‌نمایی با شرط داشتن R^*

²⁹Subtrees



شکل ۱۶.۳: ساختن درخت اتصالات T

Mutations									
Cells	□	○	△	☆	◊	▽	○	□	○
	0	0	0	0	0	0	1		
	0	1	0	0	0	0	1		
	0	1	1	0	0	0	1		
	0	1	1	0	1	0	0		
	0	1	1	0	1	1	0		
	0	1	1	0	1	1	0		
	0	1	1	0	1	1	1		
	0	1	0	0	1	0	0		
	1	1	0	0	1	0	0		
	0	0	0	1	0	0	1		
	0	0	0	1	0	0	1		

شکل ۱۷.۳: ماتریس جهش‌ها B^* با بیشینه درست‌نمایی

در اینجا $I(T)$ مجموعه نودهای داخلی درخت T را بیان می‌کند.

$$y = xxxxxxxxxxxxxxxxxx$$

مرحله اول:

$$y = xxxxxxxxxxxxxxxxxx$$

در اینجا با فرض داشتن R راست نمایی محاسبه خواهد شد و R^* با شمارش حالت جهش‌های معتبر برای هر موقعیت مکانی جهش a محاسبه شده و سپس بیشینه راست نمایی بالا حساب خواهد شد.

مرحله دوم:

یافتن زیردرخت‌ها پایش شده:

تعریف ماتریس سه‌تایی (ternary): مولفه‌های این ماتریس مقادیر ۰ و ۱ و ؟ می‌باشند.

$$y = xxxxxxxxxxxxxxxxxxxxxxxx$$

و در نهایت حل معادله برنامه‌ریزی خطی عدد صحیح زیر:

$$y = xxxxxxxxxxxxxxxxxxxxxxxx$$

منوط به شروط زیر:

با فرض اینکه M عدد ثابت و بزرگی است:

$$y = xxxxxxxxxxxxxxxxxxxxxxxx$$

نقض $F_{w,a,b}$

$G_{w,a,b}$

$$y = xxxxxxxxxxxxxxxxxxxxxxxx F_{w,a,b} \text{lr}$$

نقض $H_{w,a,b}$

$$y = xxxxxxxxxxxxxxxxxxxxxxxx F_{w,a,b} \text{lr}$$

۸.۳ الگوریتم [۹] DeepPhylo

همانطور که می‌دانیم، سرطان یک بیماری تکاملی است که با تجمع تدریجی جهش بدنی^{۳۰} در سلول‌های تومور مشخص می‌شود. رمزگشایی از تاریخچه تکاملی یک تومور، یک چالش مهم در مطالعات سرطان است و می‌تواند از جنبه‌های مهم بالینی از جمله پیشرفت تومور^{۳۱}، گشتش متاستاتیک^{۳۲} و وجود زیرکلون‌های واگرا^{۳۳} در شاخه‌های مختلف درخت فیلورژنیک تومور درک بهتری از تومور در اختیار ما بگذارد. با توجه به اهمیت

³⁰Somatic mutation

³¹Tumor progression

³²Metastatic spread

³³Divergent subclones

مسئله، تحولات سریعی در طراحی روش‌های محاسباتی اصولی برای استنباط فیلوزنی تومور وجود داشته است. بسیاری از این روش‌ها از داده‌های توالی‌یابی‌های انبوه^{۳۴} استفاده می‌کنند که DNA میلیون‌ها سلول سرطانی و طبیعی با هم یک توالی را تشکیل می‌دهند. استنباط درخت فیلوزنی با استفاده از این نوع داده‌ها، معمولاً بر مبنای دگرگونی‌های شناسایی شده^{۳۵} از بخش‌های مختلف سلول‌های سرطانی انجام می‌شود. به عنوان مثال: حذف و تغییر تعداد کپی [۸۳]^{۳۶}، دگرگونی‌های ساختاری^{۳۷} تغییر تکنولوژی‌دها [۲۷، ۴۳، ۵۱، ۶۳، ۲۴]^{۳۸}، حذف و تغییر تعداد کپی [۲۴]^{۳۹}، دگرگونی‌های ساختاری^{۳۶} [۵۷، ۲۵].

اگرچه استنباط درخت فیلوزنی با استفاده از این نوع داده مقرن به صرفه است اما رزولوشن^{۴۰} پایین داده‌های توالی‌یابی‌های انبوه یک فاکتور محدود کننده در مدل‌سازی تکامل تومور است. به طور خاص داده‌های توالی‌یابی‌های انبوه ناشی از یک نمونه تومور به طور معمول یک توپولوژی خطی را به عنوان یک راه حل بهینه در تعیین درخت فیلوزنی تومور درنظر می‌گیرد. [۲۴]

با این حال، دانستن اینکه آیا تومور شامل زیرکلون‌های واگرایی است که از طریق شاخه‌های متمایزی از فیلوزنی تومور تکامل می‌یابند، گام مهمی در جهت درک بهتر پیشرفت تومور و بهبود طرح درمانی است. تحولات اخیر تکنولوژی، محققان را قادر به انجام آزمایش‌های توالی‌یابی تک سلولی کرده است، جایی که DNA از یک سلول استخراج، تکثیر و توالی‌یابی می‌شود. توالی‌یابی تک سلولی، داده‌هایی با رزولوشن بالا برای مطالعه تکامل تومور با جزئیات زیاد را فراهم می‌کند، به عنوان مثال، امکان شناسایی توپولوژی شاخه‌ای با اطمینان بالا یا حل مشکل کلی استنباط کامل تاریخ تکامل تومور را فراهم می‌کند، حتی زمانی که تمام سلول‌های تک توالی که از یک نمونه بیوپستی^{۴۱} توموری استخراج شده باشد. روش‌های متعددی برای استنباط تاریخچه تکاملی تومور از طریق توالی‌یابی تک سلولی وجود دارد که از مهمترین آنها می‌توان به موارد زیر اشاره کرد:

- رویکردهای مبتنی بر آمار و احتمالات که از فرض مکان‌های بی‌نهایت استفاده می‌کنند. مثل الگوریتم OncoNEM^{۴۲} و IrSCITE^{۴۳}.

- رویکردهایی که از فرض مکان‌های بی‌نهایت استفاده نمی‌کنند و فرض را بر این می‌گذارند که تخطی‌های در شکل‌گیری درخت تکاملی فیلوزنی تا یک مقدار خطأ مشخص وجود دارد، مثل الگوریتم SiFit^{۴۴}.

³⁴Bulk sequencing data

³⁵Detected variants

³⁶Structural variant

³⁷Resolution

³⁸Biopsy

به تازگی الگوریتم‌هایی مثل SPhyR که از یک رویکرد بهینه‌سازی ترکیبی مبتنی بر زوجیت دولو^{۳۹} استفاده می‌کنند یا الگوریتم SiCloneFit که بهینه یافته الگوریتم SiFit می‌باشد، ارائه شده است. [۸۲، ۲۶]

شایان ذکر است که روش‌های همچون PhISCS-BnB، که از روش‌های بهینه‌سازی بر مبنای شاخه-مرز^{۴۰} استفاده می‌کنند، و یا روش‌هایی مثل ScisTree، که بر مبنای اتصال اکتشافی همسایگی^{۴۱} عمل می‌کند، به منظور بهبود زمان محاسباتی استنباط درخت فیلوزنی تومور ارائه شده‌اند. [۸۱، ۶۰]

در حالتی که هم داده‌های توالی‌یابی‌های انبو و هم داده‌های توالی‌یابی تک سلولی موجود باشد می‌توان تقریب دقیق‌تری از درخت فیلوزنی تومور بدست آورد. [۵۲، ۴۹]

همانطور که در بالا خلاصه شد، روش‌های موجود برای بازسازی فیلوزنی تومور با استفاده از داده‌های توالی‌یابی تک سلولی محدودیت‌های مهمی دارند. اولاً^{۴۲}، بسیاری از این روش‌ها، فرض مکان‌های بی‌نهایت را به کار می‌گیرند (حتی در مواقعي که شرایطی برای خطای محدود^{۴۳} و افزایش همزمان جهش‌ها^{۴۴} در نظر گرفته شود) و سطح نویز یکنواختی را در نظر می‌گیرند (منفی کاذب و همچنین نرخ مثبت کاذب) هر دو این محدودیت‌ها، با پیشرفت درک ما از تکامل تومور و فناوری توالی‌یابی تک سلولی تغییر می‌کند. مهمتر از همه، هدف از این روش‌ها استنباط محتمل‌ترین درخت فیلوزنی توموری است و برای حذف نویز (به دلیل مثال، ترک آلل یا پوشش توالی کم^{۴۵}) از روش‌های همچون بیشینه درست‌نمایی یا حداقل زوجیت^{۴۶} استفاده می‌کنند. به بیان دیگر این روش‌ها قصد دارند تا یک مساله پارامتری از مرتبه n را حل کنند ولی بدلیل عدم مقیاس‌بندی داده‌های توالی‌یابی تک سلولی به مرتبه‌های بزرگتر، در حل دقیق این مساله ناتوان هستند. حتی وقتی هدف این است که به جای بازسازی کامل درخت فیلوزنی تومور، فقط ویژگی‌های اساسی توپولوژی فیلوزنی تومور را استنباط کنیم، این روش‌ها نمی‌توانند به راحتی داده‌های توالی‌یابی تک سلولی شامل چند صد جهش و سلول را کنترل کنند. در نتیجه، تکییک‌های سریع برای استنباط ویژگی‌های کلیدی فیلوزنی تومور، به عنوان مثال، مواردی که می‌توانند توپولوژی‌های شاخه‌ای را از هم تفکیک کنند، به ویژه برای مجموعه داده‌های توالی‌یابی تک سلولی با سطح نویز بالا از محبوبیت بیشتری برخوردار هستند. به همین منظور، بهتر است در ابتدا به این سوال پاسخ داده شود که آیا حذف نویز برای ساخت فیلوزنی کامل لازم است یا خیر. سرانجام، هر یک از ابزارهای موجود به تلاش انسانی

³⁹Dollo parsimony

⁴⁰Branch-bound

⁴¹Joining-based heuristic

⁴²Limited loss

⁴³concordant gain of mutations

⁴⁴Low sequence coverage

⁴⁵Maximum parsimony

زیادی در طراحی و اجرای الگوریتمی نیاز داشته است، زیرا هر پیشرفت تکنولوژیکی در تولید داده‌ها، توسعه روش‌های کاملاً جدید را ضروری می‌کند. بنابراین داشتن یک رویکرد محاسباتی کلی که بتواند با تغییر منطقی تکنیکی سازگار شود، صرفاً از طریق آموزش آن با داده‌های جدید، بدون نیاز به مدل‌سازی صریح مشخصات نویز، بسیار مطلوب است.

رفع این محدودیت‌ها از طریق رویکرد یادگیری ماشینی یا رویکردهای "داده محور" امکان پذیر است که مجموعه‌ای کلی از توابع را در نظر گرفته و تابعی را در نهایت انتخاب می‌کند که برآورد بهتری از مجموعه داده‌های آموزشی (دادگان واقعی یا شبیه‌سازی شده) باشد. چنین رویکردی نه تنها می‌تواند از عدم دقت در مدل‌سازی مشخصات نویز بکاهد بلکه الگوهای اساسی ضمنی را در داده‌ها یا مسئله را برای توسعه اهداف واقع بینانه‌تر شناسایی می‌کند. پیشرفت‌های اخیر در یادگیری عمیق تعمیم قابل توجهی از فرمول‌بندی‌ها را برای حل بسیاری از مشکلات نشان داده است. [۴۸، ۶۷، ۲۲]

این امکان وجود دارد که یک معماری یادگیری عمیق، زمانی که بتواند در تعداد کافی مجموعه داده آموزش را دیده باشد، بتواند در استنباط خواص متمایز از فیلوزنی‌های تومور موفق شود. در سالهای اخیر، بسیاری از برنامه‌های محاسباتی، رویکرد الگوریتمی خود را به رویکردهای داده محور تغییر داده‌اند. مانند رمزگشایی متن دست نوشته برای شناسایی رقم [۱۶] و پردازش زبان طبیعی. [۲۲]

مسائلی که در بایولوژی ساختار یافته، فرمول‌سازی هدفمند یا کمی‌سازی آنها مشکل است (مانند استنباط ساختار سه بعدی توالی پروتئینی) از روش‌های مبتنی بر یادگیری عمیق بسترین استفاده را در جهت حل مسائل خواهند کرد. [۶۶]

با این حال این مقاله، اولین مقاله استنباط درخت فیلوزنی تومور مبتنی بر رویکردهای داده محور است. در این مقاله، اولین روش‌های بازسازی فیلوزنی تومور مبتنی بر داده را برای رفع محدودیت‌های استراتژی‌های موجود ارائه شده است. نویسنده‌گان این مقاله از داده‌های توالی یابی تک سلولی در کنار شبکه‌های عصبی عمیق و یادگیری تقویتی برای استنباط ویژگی‌های تپولوژیکی فیلوزنی تومور و همچنین محتمل‌ترین سابقه تکاملی تومور استفاده شده‌است. برای رسیدن به این هدف، چندین چالش وجود داشت:

۱. شبکه عصبی در حالت ایده‌آل باید طوری طراحی شود که بتواند تعداد متفاوتی از سلول‌ها و جهش‌ها را کنترل کند. متناوباً، برای مدل‌هایی با ورودی‌هایی با اندازه ثابت، بهتر است که از دانش خود در زمینه تهیه داده استفاده شود تا داده‌ها به روشی تهیه شود تا موفقیت در پیش‌بینی‌ها را تسهیل کند.

۲. با توجه به استفاده از شبکه‌های عصبی، برای آموزش مناسب به تعداد زیادی نمونه نیاز است. متأسفانه، تعداد مجموعه داده‌های توالی‌یابی تک سلولی تومور در دسترس عموم برای آموزش مدل‌های یادگیری عمیق به اندازه کافی زیاد نیست. بنابراین، نیاز به تولید تعداد زیادی مجموعه داده شیوه‌سازی شده داده‌های توالی‌یابی تک سلولی وجود دارد.

۳. نویز و خطاهای موجود در داده‌های توالی‌یابی تک سلولی پیچیدگی بیشتری را به این مسئله می‌افزاید و چارچوب پیشنهادی یادگیری عمیق باید از نظر تحمل نویز ارزیابی شود.

۴. معماری انتخاب شده مستلزم نوع خاصی از نظارت است که ما باید قادر به تامین آن باشیم.

به منظور کاهش یا حذف نویز در ورودی "ماتریس ژنوتیپ" استخراج شده از داده‌های توالی‌یابی تک سلولی، می‌توان نظارت را به صورت مجموعه داده‌ای از ورودی‌های نویزدار به همراه با ورودی‌های بدون نویز ارائه داد. یک نظارت جایگزین و ارزان‌تر توسط مکانیزم بازخورد^{۴۶} است که تعیین می‌کند که آیا یک خروجی از شبکه عصبی با موفقیت بدون نویز شده است یا خیر. گزینه سوم توسط یکتابع هزینه ارائه می‌شود که به طور غیر مستقیم کمک به نظارت بر فرایند یادگیری تقویتی می‌کند.

در این مقاله با الهام از رویکردهای جدید یادگیری عمیق برای مسائل گوناگون مانند "الگوریتم گرادیان سیاست تقویتی" برای مساله فروشنده دوره گرد^{۴۷}[۸۰]، رویکرد NeuroSAT [۶۵] برای مساله رضایتمندی با استفاده از نظارت تکبیتی، یک چارچوب محاسباتی ایجاد شد تا همه چالش‌های فوق را به شرح زیر با موفقیت حل کند.

۱. یک رویکرد مبتنی بر یادگیری تقویتی به منظور آموزش مدلی جهت از بین بردن نویز داده‌ها بدون نیاز به استاندارد مرجع^{۴۷} به کار گرفته شد. تابع هزینه استفاده شده در این مدل یکتابع هزینه خاص برای رفع مساله از بین بردن نویز بود.

$$y = xxxxxxxxxxxxxxxxxxxxxxx$$

که در آن X ماتریس خروجی ناشی از ورودی A' است.

⁴⁶feedback

⁴⁷Gold standard

۲. داده‌های ماتریس ورودی، که از مجموعه دادگان نویزی توالی‌یابی تک سلولی استخراج شده، در کنار نرخ

نویز و موقعیت مکانی به عنوان ورودی به شبکه داده شده است. این رویکرد در مجموعه دادگانی با سایز

متفاوت همچنان کارآمد است و مستقل از جابجایی در سطر و ستون ماتریس ورودی است.

۳. یک مرحله پیش‌پردازش دیتا، به منظور به کارگیری دانش حاصل از تجربه در نظر گفته شده است تا هر

گونه عملکردی را که می‌تواند پیش‌بینی مدل را بهبود بخشد، بر روی داده‌ها اعمال گردد.

۴. داده‌های شبیه‌سازی شده ورودی مدل از طریق یک چاچوربی که راستی آزمایی شده است، توسعه یافته

است.

در نمودار ۱۸.۳، میزان دقت علمکرد شبکه در حذف نویز داده ورودی و تاثیر مرحله پیش‌پردازش بر خروجی الگوریتم را مشاهده می‌کنید. همچنین تاثیر میزان نرخ نویزی بودن داده‌ها در خروجی شبکه قابل توجه است.

تصاویر A و C میزان دقت شبکه در حذف نویز داده‌ایی را نشان می‌دهد که با نرخ نویزهای $\alpha = 0.02$

$\beta = 0.1$ نمونه‌برداری شده‌اند اما تصاویر B و C میزان دقت شبکه در حذف نویز داده‌ایی را نشان می‌دهد که

با نرخ‌های کاذب مثبت $\alpha = 0.00004$ و کاذب منفی $\beta = 0.002$ نمونه‌برداری شده است.

همچنین در جدول ۲.۳ تاثیر مرحله پیش‌پردازش دیتا در دقت خروجی مدل در حذف نویز از دیتا را مشاهده

می‌کنید که میزان دقت حذف نویز بهبود قابل قبولی داشته است.

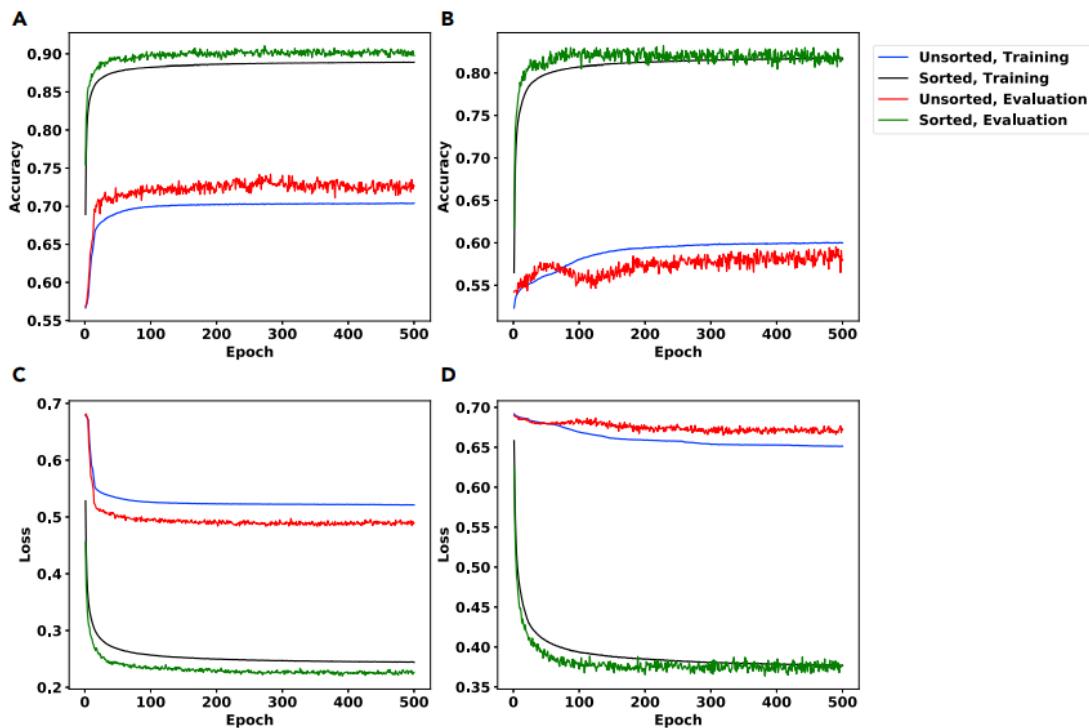
Input MAtrix Size	A	B	Unsorted Acc.	Sorted Acc.
10*10	0.002	0.1	72	90
10*10	$4 * 10^{-4}$	0.02	60	81
25*25	$3.2 * 10^{-4}$	0.016	50	77
25*25	$6.4 * 10^{-4}$	0.0032	52	65

fffffffffffff 3.2:

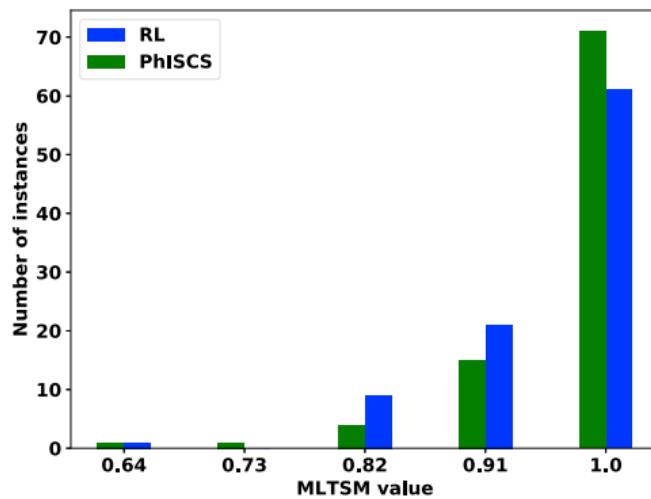
در نهایت مقایسه بین عملکرد الگوریتم پیشنهادی در این مقاله و الگوریتم PhISCS با استفاده از معیار شباهت MLTSM84 انجام شد که نتیجه این مقایسه در شکل ۱۹.۳ آمده است. همانطور که در شکل

مشهود است عملکرد الگوریتم پیشنهادی در میزان شباهت‌های مشابه، تعداد استنباط‌های بیشتری از فیلوزنی

تومور را شامل می‌شود.

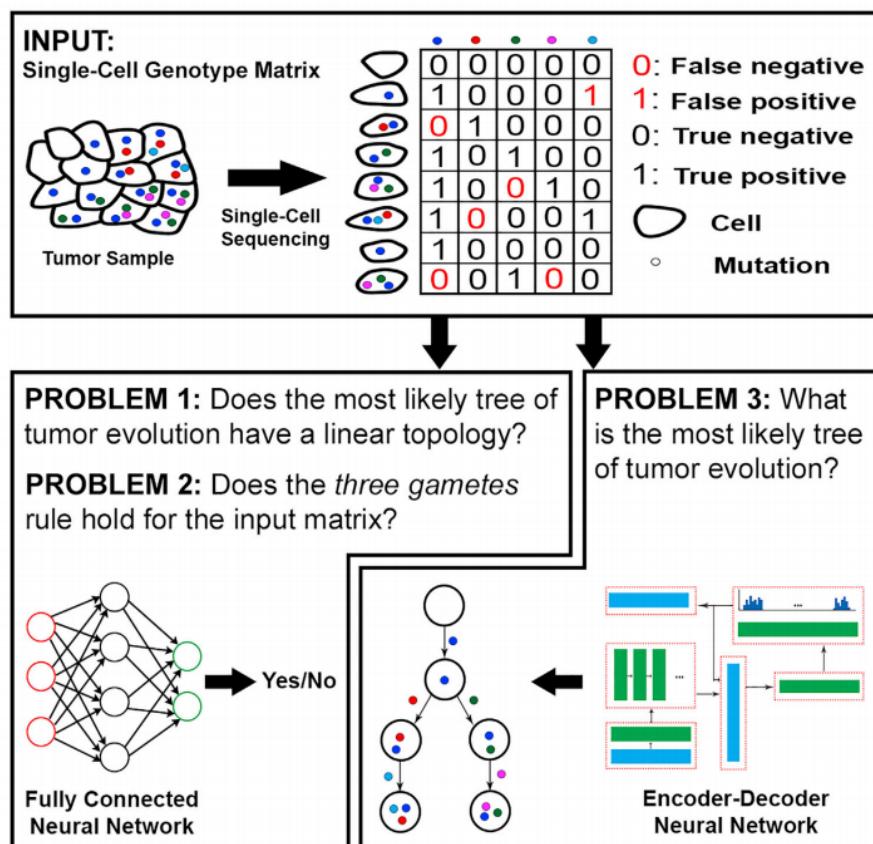


شکل ۱۸.۳: عکس



شکل ۱۹.۳: عکس

خلاصه‌ای از سازکار به کار رفته در این مقاله در شکل ۲۰.۳ آمده است:



شکل ۲۰.۳: عرضه

۹.۳ جمع‌بندی

جهش‌های سومایتک تومورها در تمام مقیاس‌های ژنومی، از دگرگونی‌های تک‌هسته‌ای (SNV) تا جهش‌های حذف و تغییر تعداد کپی (CNA) وجود دارد. تا به امروز، بیشتر روش‌های ساخت فیلوزنی توموری از داده‌های توالی‌یابی تک‌سلولی DNA فقط از دگرگونی‌های تک‌هسته‌ای استفاده می‌کردند. [۲۶، ۵۰، ۵۴، ۶۸]

جهش‌های حذف و تغییر تعداد کپی و در نتیجه اطلاعات مهم استنباط فیلوزنیک تومور را نادیده می‌گرفتند.

وجود ناهمگنی‌های درون توموری باعث ناکارآمدی درمان‌های دارویی تومور می‌شود زیرا که هر یک از این روش‌های درمانی به طور موثر فقط بر روی تعداد محدودی از کلون‌های توموری اثر می‌گذارند و همه این زیرنوایی را تحت تاثیر قرار نمی‌دهند. با مطالعه بر روی تومورهای مختلف این امکان حاصل می‌شود تا الگوهای درون توموری بهتر شناخته شوند و درمان دارایی تومور کارآمدتر و بهینه‌تر از گذشته صورت پذیرد. مطالعه بر

روی داده‌های توالی‌یابی تک‌سلولی یکی از زمینه‌های تحقیقاتی است که می‌تواند منجر به افزایش دانش از نحوه شکل‌گیری و تکامل تومور شود. پیدا کردن سیر زمانی تکامل تومور با استفاده از داده‌های توالی‌یابی تک‌سلولی چالشی است که اخیراً مورد توجه قرار گرفته است که در جدول زیر خلاصه‌ای از روش‌هایی که در این فصل مورد بررسی قرار گرفته‌اند، مشاهده می‌شود.

جدول ۳.۳: Comparison

Method	Dataset	Algorithm	Output	Evaluation Method	Limitation
Kim & Simon approach	Thrombocytopenia Essential (TE)	Minimal spanning tree of the Edmonds' algorithm	Phylogenetic tree	Leave one out cross validation	high computational time and excluding uncertainty dataset error
BitPhylogeny	JAK2 negative myeloproliferative	TSSB, MCMC	Evolutionary clonal tree	V measure comparison with K-Centroids and Hierarchical Clustering	High computational time, infinite sites assumption and homozigous differentiation
SCITE	JAK2 negative myeloproliferative, clear cell and renal cell carcinoma, estrogen-receptor positive breast cancer	Maximum bayesian MCMC likelihood,	Phylogenetic tree	Better performance in real dataset in comparison with biphylogeny algorithm	Infinite sites assumption
ONCONEM	Muscle invasive bladder transitional cell carcinoma	Neighbor joining, MCMC	Phylogenetic tree, evolutionary clonal tree	Score function extracted from nested model	Infinite sites assumption, homozigous heterozygous differentiation
SASC	Muscle invasive bladder transitional cell carcinoma, Thrombocytopenia Essential (TE) Robust approach based on scDNA-seq data from a metastatic colorectal cancer patient	Simulated annealing	Phylogenetic tree	Better performance in real dataset in comparison with SCITE algorithm	Limited mutation assumption
SCARLET	Acute Lymphoblastic Leukemia, TNBC dataset	Loss-Supported Phylogenetic Model	Phylogenetic tree	Mutation matrix error and pairwise ancestral relationship error	Mutation loss due to the dollo assumption
DeepPhylo		Critic-actor reinforcement learning	Phylogenetic tree	Accuracy, maximum likelihood	Fixed input dimension, lack of empirical experiment

فصل ۴

روش پیشنهادی

۱.۴ مقدمه

پس از آشنایی با روش‌های پیشین که برای حل مسئله مشابه مورد استفاده قرار گرفته‌اند، حال می‌توانیم به معرفی و تشریح روش پیشنهادی خود برای حل مسئله پیش رو پردازیم. در این فصل ابتدا داده‌های ورودی مسئله را همراه با فرضیات در نظر گرفته شده بیان می‌کنیم و پس از آن روش پیشنهاد خود را بیان خواهیم نمود. این روش با الهام از ۳ روش قبلی متفاوت تنظیم شده است. در ابتدا پایه و بنیان آن به یکی از رویکردهای پیشین نزدیک‌تر است که با تغییری از جنس روش‌های نوین در مراحل میانی به یک روش جدید می‌رسیم که به علت افزایش سرعت همگرایی می‌توان فرض و داده‌های جدیدی را از طریق حذف و تغییر تعداد کپی به آن افزود و پاسخ گرفت که این عمل با بهره‌گیری از رویکردی جدید در حوزه یادگیری ماشین همراه است که به کمک یادگیری تقویتی به حل مسئله مورد نظر می‌بردارد.

۲.۴ معرفی دادگان ورودی

قبل از وارد شدن به بخش روش‌های پیشنهادی نیاز است تا دادگان ورودی را مشخص و معرفی نماییم. دادگان ورودی در این پایان‌نامه همگی به صورت فایل‌های خام اسکی^۱ هستند که حاوی اطلاعات جهش‌های ماتریسی

¹Ascii

ژن-سلول (SNV) و اطلاعات مربوط به حذف و تغییر تعداد کپی هستند.

در ادامه جدول ۱.۴ را برای معرفی اندیس‌های بکار گرفته شده در روابط مربوط به روش پیشنهادی اول معرفی می‌نماییم.

جدول ۱.۴: اندیس‌های به کار رفته در روابط روش پیشنهادی اول

ماتریس داده نویزی در دسترس که مقادیر ۰ و ۱ در آن قرار دارد	D
ماتریس داده حقیقی بدون نویز که به دنبال آن هستیم	E
درخت فیلوزنی جهش‌ها	T
بردار انتصابات	σ
بردار پذیرش فقدان	\varnothing
ماتریس متناظر درخت	X_T
تعداد سلول‌های نمونه	N
تعداد جهش‌ها	M
مجموعه سلول‌های متمایز از هم	N
مجموعه جهش‌های متمایز از هم	M
مجموعه جهش‌های با پتانسیل حذف	L
نرخ خطای مثبت کاذب	α
نرخ خطای منفی کاذب	β

۳.۴ روش پیشنهادی برای مدیریت داده‌های از دست رفته

در ادامه این بخش به معرفی روش‌های پیشنهادی پرداخته خواهد شد اما در ابتدا به دلیل وجود داده‌های از دست رفته در پایگاه‌داده‌های مورد استفاده لازم است تا به بررسی و ارائه رویکردی برای حل این مشکل پرداخته شود و در ادامه پس از معرفی روش پیشنهادی برای مدیریت این داده‌های از دست رفته، هر کدام از روش‌های پیشنهادی به تفضیل شرح داده شود.

همان‌گونه که در داده‌های حقیقی مشاهده شد در پایگاه داده‌های حقیقی ما با اطلاعات از دست رفته مواجه هستیم و به همین دلیل نیز سعی کردیم تا در پایگاه داده مجازی تولید شده نیز به مشابه داده‌های حقیقی، شامل اطلاعات از دست رفته باشد. در این بخش به رویکرد روش محاسبه استاتیک برای مدیریت این داده‌های از دست

رفته می‌پردازیم و در بخش بعد به معنی روشنی برای بدست آوردن درخت فیلوزنی پرداخته خواهد شد. همان‌گونه که در ادامه بررسی خواهد شد، این اطلاعات از دست رفته در پایگاه داده‌های مختلف نرخ‌های متفاوتی دارد که تاثیر این تغییرات نیز در روشنی پیشنهادی بررسی خواهد شد.

۱۰.۳.۴ روش محاسبه استاتیک

در این روش قصد داریم تا به یکباره بتوانیم مقادیر مناسب برای داده‌هایی که از دست رفته‌اند را تخمین بزنیم. در این روش باید توجه شود که ما لزوماً به دنبال جایگذاری مقدار از دست رفته با مقدار درست واقعی نیستیم. اگرچه چنین بیانی در نگاه اول ممکن است تعجب‌آور باشد اما با دقت بیشتر متوجه خواهیم شد که ما در آینده برای خطاهای موجود در پایگاه داده مدل‌سازی‌های محدودی داریم. مدل‌هایی که بهترین آن‌ها نیز ممکن است با واقعیت نویز افزوده شده به دادگان متفاوت باشد. در نتیجه اگر مطمئن بودیم که تمام داده‌هایی که موجود می‌باشند بدون خطأ هستند در آن صورت ما نیز به دنبال یافتن جایگذاری با مقدار واقعی بودیم اما در حال حاضر که درصدی از داده‌های در دسترس خود همراه با خطأ می‌باشند، ما به دنبال جایگذاری‌ای هستیم که بتواند در مجموع با مدل‌سازی خطایی که در نظر می‌گیریم بیشترین سازگاری را داشته باشد کما اینکه ممکن است در حقیقت جایگزاري اشتباхи انعام داده باشیم. حال با توجه به توضیحی که بیان شد به تشریح این روش می‌پردازیم.

با توجه به فرض مدل مکان‌های بی‌نهایت می‌دانیم که جهش‌های اتفاق افتاده در والد در تمامی نسل‌های آینده باقی خواهد ماند. بنابرین اگر تمامی جهش‌های نمونه (سلول) a در نمونه‌ای دیگر مانند b قرار داشته باشد، بنابرین می‌توان نتیجه گرفت که a یکی از اجداد b خواهد بود. همین فرضیه هسته اصلی روش پیشنهادی در نظر گرفته شده را تشکیل می‌دهد. بنابرین اگر جهش i در سلول a از دست رفته است، با توجه به اینکه آن جهش در سلول b چه وضعیتی دارد می‌توان تصمیم‌گیری کرد. اگر $b(i) = 0$ باشد، در این صورت $a(i)$ حتماً باید 0 باشد و گرنه فرض اولیه مدل مکان‌های بی‌نهایت نقض خواهد شد. اما اگر $b(i) = 1$ باشد، آنگاه نتیجه خاصی نمی‌توان گرفت و باید به دنبال نمونه والد a یعنی نمونه d باشیم. حال اگر $d(i) = 1$ باشد، آنگاه $a(i)$ حتماً باید 1 باشد. اما اگر $d(i) = 0$ باشد آنگاه انتخاب هر مقداری برای $a(i)$ تقریباً آزاد خواهد بود زیرا با فرض اولیه تناقضی ندارد و اینکه ساختار فیلوزنی را تغییر نمی‌دهد. اما از آنجایی که خود داده‌های در دسترس شامل خطأ می‌باشند و هر نمونه‌ای که حاوی اطلاعات از دست رفته است لزوماً یک نواده یا یک والد ندارد، مجموعه‌ای از سلول‌های فرزند

یا والد خواهند بود که متناسب با پارامترهای خطابی که در نظر می‌گیریم و فاصله‌زنی‌ای که دارند می‌توانند در تصمیم‌گیری تاثیرگزار باشند. صورت دقیق‌تر توضیحات داده شده را می‌توان به صورت فرمولی که در ادامه آمده است به نمایش درآورد.

در ابتدا تابعی به نام $F_s(D_{ij})$ تعریف می‌کنیم که به نوعی با توجه به ارزشی که به سلول‌های نواده شده از سلول j می‌دهد سعی دارد تا اطمینان ۰ بودن داده از دست رفته D_{ij} را بیان کند.

برای محاسبه این تابع می‌دانیم که ابتدا سلول‌های مختلف با توجه به احتمال نواده بودنشان باید رتبه‌بندی شوند و وزن بگیرند. پس از آن هر سلول متناسب با ارزش تاثیرگزاری خود می‌تواند در مورد جایگاه جهش \circ برای سلول j نظر دهد.

$$F_s(D_{ij}) = \sum_{n \in \mathcal{N}} (1 - D_{mj}) \prod_{m=1}^M W(D_{mn}, D_{mj}) \quad (1.4)$$

در فرمول ۱.۴ مجموعه \mathcal{N} برابر با مجموعه سلول‌های متمایز از هم است. زیرا که در بسیاری از پایگاه‌داده‌ها از یک نمونه سلول ممکن است چندین نمونه وجود داشته باشد که وجود آن‌ها باعث بایس در محاسبات ما خواهد شد. همچنین تابع $W_s(c, p)$ به ارزش‌دهی جهش c در برابر p به عنوان نواده بودن می‌پردازد که در فرمول ۲.۴ تعریف شده است.

$$W(c, p) = \begin{cases} 1 & \text{if } c = 1, p = 1 \\ 1 - \xi & \text{if } c = 1, p = \circ \\ 0 & \text{if } c = \circ, p = 1 \\ 1 & \text{if } c = \circ, p = \circ \end{cases} \quad (2.4)$$

مقدار ξ عددی بین $(1, 0)$ است که پارامتری در جهت میزان ارزش‌دهی به نوادگان با فواصل مختلف می‌باشد. هرچه این عدد بزرگتر باشد به معنی کم ارزش‌تر شدن نوادگان با فواصل بیشتر است و بلافاصله.

به همین صورت برای اولاد سلول j نیز می‌توان مشابه حالت قبل عمل کرد که روابط آن به صورت فرمول ۳.۴

خواهد شد.

$$F_a(D_{ij}) = \sum_{n \in \mathcal{N}} D_{mj} \prod_{m=1}^M W(D_{mj}, D_{mn}) \quad (3.4)$$

حال دو نکته در استفاده از روابط بالا باقی خواهد ماند.

نکته اول وجود داده‌های دیگر از دست رفته در محاسبه توابع است که به دو صورت می‌توان با آن‌ها برخورد نمود. رویکرد اول این است که در آنجاییگاه‌هایی از محاسبه آن خود داری شود و رویکرد دوم استفاده از از مقدار ۵٪ را فراوانی نسبی آن جهش در محاسبات است که ما رویکرد اول را در این گزارش استفاده خواهیم کرد.

نکته دوم وجود خطأ در داده‌های است. برای مدیریت این مشکل می‌توان با مدل‌سازی خطأ که به صورت فرمول ۴.۴ بیان می‌شود، برخورد کرد.

$$\begin{aligned} P(D_{ij} = 1 | E_{ij} = 0) &= \alpha, & P(D_{ij} = 0 | E_{ij} = 0) &= 1 - \alpha \\ P(D_{ij} = 0 | E_{ij} = 1) &= \beta, & P(D_{ij} = 1 | E_{ij} = 1) &= 1 - \beta \end{aligned} \quad (4.4)$$

پس از تعریف مدل‌سازی خطأ می‌توان روابط قبلی را مجدداً به صورتی که در ادامه آمده است بازنویسی کرد.

$$W_e(c, p) = \sum_{i,j \in \{0,1\}} P(c|E_c = i) P(p|E_p = j) W(i, j) \quad (5.4)$$

که در این صورت توابع F_a و F_p نیز به صورت زیر همراه با مدل‌سازی خطأ بازتعریف خواهند شد.

$$\begin{aligned} \hat{F}_s(D_{ij}) &= \sum_{n \in \mathcal{N}} [1 - D_{mj}(1 - \alpha)] \prod_{m=1}^M W_e(D_{mn}, D_{mj}) \\ \hat{F}_a(D_{ij}) &= \sum_{n \in \mathcal{N}} D_{mj}(1 - \beta) \prod_{m=1}^M W_e(D_{mj}, D_{mn}) \end{aligned} \quad (6.4)$$

حال پس از محاسبه مقادیر \hat{F}_s و \hat{F}_a می‌توان در مورد داده نامعلوم D_{ij} به صورت فرمول ۷.۴ تصمیم گرفت.

$$D_{ij} = \begin{cases} 0 & \text{if } \hat{F}_s \geq \hat{F}_a \\ 1 & \text{if } \hat{F}_s < \hat{F}_a \end{cases} \quad (7.4)$$

همچنین با کمی دقت در فرمول‌بندی انجام شده اگر برای تمام j, i ‌های ماتریس D این مقادیر توابع \hat{F} محاسبه شوند، خود می‌توانند معیاری برای ارزیابی پایگاهداده در دسترس و احتمال درستی فرض مدل مکان‌های بی‌نهایت باشند.

۲.۳.۴ تصادفی

پر کردن کاملاً تصادفی میس‌ها. در این روش به صورت تصادفی مقادیر از دست رفته را مقدار دهی می‌کنیم. تنها نکته‌ای که در این روش وجود دارد این است که نباید این پرکردن تصادفی داده‌های از دست رفته باعث شود تا پارامترهای مدل‌سازی ای که از قبیل در نظر گرفته بودیم با این روش نادقيق شوند.

۴.۴ روش پیشنهادی

در این روش ما بر حسب بهتر کردن یک پاسخی که از پیش داشتیم به دنبال رسیدن به بهترین پاسخ ممکن در طی تکرار پشت سر هم هستیم. برای مشخص شدن نحوه کارکرد روش پیشنهادی در مراحلی که در ادامه بیان

خواهد شد به عنوان مثال یک ماتریس

$$D = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad (8.4)$$

را به عنوان ورودی مساله به همراه پارامترهای α و β در نظر بگیرید. (برای راحتی کار فرض کرده‌ایم که داده از دست رفته در D نداریم.)

۱.۴.۴ پیش‌پردازش

قبل از شروع باید بر روی داده‌ها یک پیش‌پردازش اعمال کنیم که وابسته به سیاست درنظر گفته شده می‌تواند باعث تغییر در پاسخ نهایی نیز شود. به این منظور داده‌ایی که miss شده‌اند با یکی از دو روشی که معرفی شد تخمین رده می‌شوند و برای ورود به مرحله بعد آماده می‌شوند.

۲.۴.۴ اولین پاسخ (درخت تصادفی)

همان‌گونه که از قبل می‌دانستیم خروجی نهایی ما برابر با درختی خواهد بود که نودهای آن برابر با جهش‌های ماتریس ورودی ما و برگ‌های آن برابر با نمونه‌های مشاهده شده خواهند بود. در روش پیشنهادی اول ما به دنبال بهتر کردن این درخت به عنوان پاسخ هستیم. از این رو پایه این روش پیشنهادی اول بر مبنای بهتر کردن پاسخ فعلی بنا نهاده شده است. در نتیجه ما همواره پاسخی به عنوان جواب نهایی داریم که تلاش خواهیم نمود تا با استفاده از ابزارهایی بتوانیم ابا ایجاد تغییری در این پاسخ به پاسخی جدید برسیم که قابل مقایسه با پاسخ فعلی برای انجام مراحل بعدی باشد.

با توجه به توضیحاتی که داده شد ما برای شروع الگوریتم پیشنهادی اول خود نیاز به یک پاسخ داریم. این پاسخ

که درخت فیلوزنی هست با توجه پارامترهای ورودی و انتخاب یک نود (زن) $root$ به عنوان ریشه این درخت به صورت زیر حاصل می‌شود.

$$\begin{aligned} \mathcal{M} &= \{1 \dots M\} \\ \hat{B}_{T\setminus} &= [R_1(\mathcal{M} - |1|), R_2(\mathcal{M} - |2|), \dots, R_{root}(\{\}), \dots, R_M(\mathcal{M} - |M|)] \end{aligned} \quad (9.4)$$

که در این رابطه \mathcal{M} برابر با مجموعه تمامی جهش‌های متمایز از شماره ۱ تا M است و \hat{B} مشخص کننده نود پدر در درخت برای جهش i ام در این لیست خود است که توسطتابع $(X) R_i(X)$ به صورت کاملاً یکنواخت^۲ از اعضای مجموعه X انتخاب می‌شود.

با توجه به مثالی که در رابطه ۸.۴ زده شد فرض کنید مقدار بردار \hat{B} با ریشه ۲ $= root = 2$ به صورت رابطه ۱۰.۴ شود.

$$\hat{B} = [2, 1, -, 0] \quad (10.4)$$

که درخت شکل ۱.۴ را نتیجه می‌دهد.

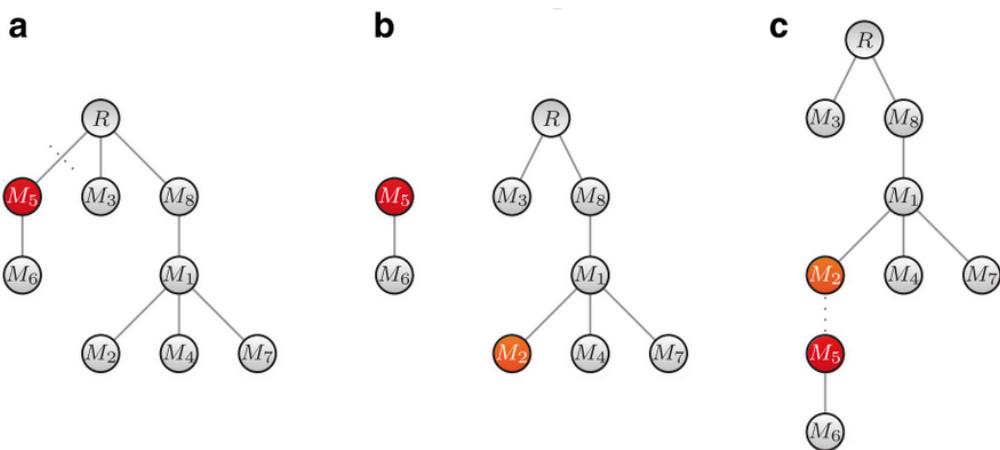
شکل ۱.۴: درخت تصادفی اول

۳.۴.۴ پاسخی جدید

تا به اینجا ما یک درخت فیلوزنی به عنوان پاسخ داریم که در این بخش می‌خواهیم با انجام تغییراتی بر روی آن به یک پاسخ جدید برسیم تا در گام‌های بعدی بتوانیم با مقایسه آن‌ها تصمیمات لازم را برای ادامه الگوریتم بگیریم. به همین منظور تقریباً مشابه با روش [۱۸] به صورت هرس و اتصال دوباره^۳ قصد داریم تا درخت پاسخ فعلی را برای رسیدن به یک پاسخ دیگر تغییر دهیم. در شکل ۲.۴ مثالی از این روش آورده شده است. در این شکل مطابق با قسمت a یکی از نودهای درخت (به جز ریشه) انتخاب می‌شود (در اینجا نود M_5) و اتصال از

²Uniform

³Prune and reattach



شکل ۲.۴: نحوه انجام کار روشن هرس و اتصال دوباره [۱۸].

پدرش قطع می‌شود. در این حالت به دو درخت مشابه با شکل b می‌رسیم. حال در درخت باقی مانده یک نod دیگر (M_2) به عنوان پدری جدید انتخاب می‌شود تا با این تغییر به درخت جدید شکل c بررسیم. در نتیجه با این تکنیک می‌توان به پاسخ‌های جدید رسید اما سوالی که باقی ماند این است که این دونود چگونه باید انتخاب شوند؟ این دو انتخاب به صورت هوشمند توسط دو شبکه عمیق گرفته می‌شود. این شبکه‌ها که در اصل شبکه‌های یادگیری تقویتی عمیق هستند، از پیش برای این منظور آموزش داده شده‌اند. شبکه اول برای انتخاب محل برش (هرس) مورد استفاده قرار می‌گیرد و شبکه دوم با دریافت خروجی ×xxxxxx دو درخت راهبر-پیرو^۴ مکان‌های مناسب برای بازاتصال را ارزش‌گزاری می‌کند. این دو شبکه در قسمت‌های ۷.۴.۴ و ۸.۴.۴ به تفصیل شرح داده شده‌اند.

۴.۴.۴ مقایسه و ارزیابی پاسخ‌ها

پس از اینکه از پاسخ فعلی به یک پاسخ جدید رسیدیم حال می‌توان کیفیت این دو پاسخ را باهم مقایسه کرد و پس از آن با توجه به امتیاز دو پاسخ در مورد پذیرش یا عدم پذیرش پاسخ جدید در برابر پاسخ فعلی تصمیم گرفت. این فرآیند شامل دو بخش اصلی است که در دوزیربخشی که در ادامه آمده است بیان شده‌اند.

⁴Master-slave

۱۰.۴.۴ تبدیل درخت پاسخ به ماتریس

برای ادامه روش پیشنهادی و مقایسات لازم است تا درخت پاسخ را به ماتریس X تبدیل کنیم که قابل بررسی با داده‌های مشاهده شده D باشد. ماتریس X مشابه با ماتریس D مشکل از مقادیر \circ و ۱ خواهد بود که به عنوان مثال $\text{۱} = X_{i,j}$ به این معنی است که طبق درخت T در سلول i جهش j مشاهده نشده است. ما هر درخت T را می‌توانیم با مقادیر مختلفی از σ و \circ مزین کنیم و به ماتریس‌های مختلفی بررسیم. اما در نهایت مهمترین پارامترها که بیشترین امتیاز را برای درخت ما بوجود می‌آورند مطلوب ما خواهند بود و ماتریس متناظر با آن حالت را X می‌نامیم و به مراحل بعدی برای محاسبات انتقال می‌دهیم. در نتیجه کار ما در این بخش این خواهد بود که به ازای درخت دلخواه T بتوانیم بهترین σ و \circ را بدست آوریم و از روی آن‌ها ماتریس متناظر X را بدست آوریم. از پیش با بررسی پروفایل‌های شماره کپی^۵ در بخش ۵.۴.۴ به \mathcal{L} رسیده‌ایم که مشخص می‌کند چه جهش‌هایی پتانسیل حذف را دارند و در این بخش زمان استفاده از این اطلاعات است. در ادامه ماتریس B را به صورت رابطه ۱۱.۴ تعریف می‌کنیم که برای انتخاب بهینه σ مورد استفاده قرار خواهد گرفت.

$$B_{i,j} = \begin{cases} \text{۱} & \text{if } j = i \text{ یا } i \text{ یکی از نوادگان } j \text{ باشد که در } \mathcal{L} \text{ نباشد} \\ x & \text{if } i \text{ یکی از نوادگان } j \text{ باشد و همچنین در } \mathcal{L} \text{ باشد} \\ \circ & \text{در صورتی که دو مورد بالایی نباشد} \end{cases} \quad (۱۱.۴)$$

در رابطه بیان شده x به معنای این است که هم می‌تواند مقدار \circ و هم مقدار ۱ را داشته باشد. حال می‌توانیم برای هر سلول (نمونه) c_i در ماتریس مشاهده شده D امتیاز اتصال را در هر قسمت از درخت T حساب می‌کنیم که از طریق رابطه ۱۲.۴ بدست می‌آید.

$$S(c_i, T, k) = \prod_{j=0}^M P(D_{i,j} | B_{j,k}) \quad (۱۲.۴)$$

در این رابطه $S(c_i, T, k)$ برابر امتیاز اتصال نمونه i در درخت T در مکان زن (جهش) k است. ناگفته نماند که،

$$P(D = \text{۱} | B = x) = 1 - \beta, \quad P(D = \circ | B = x) = 1 - \alpha \quad (۱۳.۴)$$

^۵Copy number profile

بنابرین به ازای هر x ما دو حالت را می‌توانیم داشته باشیم که آن‌ها همان پذیرش یا عدم پذیرش حذف جهش‌های در مجموعه \mathcal{L} است. برای اینکه بهترین σ را داشته باشیم باید بتوانیم این امتیازاتی که با پذیرش‌های مختلف x بدست می‌آیند را به ازای تمام نمونه‌های در دسترس ثبت و بررسی کنیم. برای این منظور رابطه [۱۱.۴](#) را به صورت رابطه [۱۴.۴](#) بازنویسی می‌کنیم.

$$B_{i,j} = \begin{cases} 1 & \text{if } \text{اگر } \mathbf{z}_i \text{ یکی از نوادگان } \mathbf{z}_j \text{ باشد که در } \mathcal{L} \text{ نباشد} \\ x_i^{\text{dist}(i,j)} & \text{if } \text{اگر } \mathbf{z}_i \text{ یکی از نوادگان } \mathbf{z}_j \text{ باشد و همچنین در } \mathcal{L} \text{ باشد} \\ 0 & \text{if } \text{در صورتی که دو مورد قبلی نباشد} \end{cases} \quad (14.4)$$

همان‌گونه که در این رابطه بیان شده همچنان مقادیر نامشخص وجود دارد. برای مشخص کردن این مقادیر نامشخص از σ استفاده می‌کنیم. σ یک لیست به طول تعداد ژن‌هایی است که در مجموعه \mathcal{L} قرار دارند. نتیجه اعمال σ بر B ماتریس A را نتیجه خواهد داد که به صورت زیر تعریف می‌شود.

$$A_{i,j} = \begin{cases} 1 & \text{if } \sigma_i < \text{dist}(i,j) \text{ یا } B_{i,j} = 1 \\ 0 & \text{if } \text{شرط بالا درست نباشد} \end{cases} \quad (15.4)$$

این مقادیر نامشخص که در اتصال به ژن \mathbf{z}_i و نوادگان آن در درخت مشخص شده‌اند با توجه به فاصله تعیین شده از این ژن \mathbf{z}_i برای نوادگان حذف خواهد شد که این فاصله در σ_i مشخص شده است. پس حال با تعیین مقادیر بردار σ می‌توانیم بهترین B نامعلوم را به A معلوم تبدیل کنیم. به رابطه [۱۲.۴](#) توجه کنید. این رابطه امتیاز اتصال نمونه \mathbf{z}_i را به مکان k در درخت بیان می‌کند. ما به دنبال محلی هستیم که ضمیمه کردن نمونه به آن محل بالاترین امتیاز را بدست آورد. بنابرین σ را به صورت یک بردار به طول N (تعداد نمونه‌ها) تعریف می‌کنیم به طوری که شماره اندیس i در آن متناظر با \mathbf{z}_i نمونه در ماتریس D باشد و مقداری که در آن خانه از σ قرار می‌گیرد برابر با شماره یکی از ستون‌های ماتریس A باشد که نشان‌دهنده بهترین محلی است که در درخت T می‌تواند به آن ضمیمه شود. حال می‌توانیم جایگاه هر اتصال به درخت را که بالاترین امتیاز را به ارمنغان می‌آورد مشخص کنیم

و پس از آن به تبدیل درخت T به ماتریس X بپردازیم.

$$\begin{aligned} S(c_i, T) &= \max_{k \in \{0, \dots, M\}} S(c_i, T, k) \\ &= \max_{j^*} \left(\prod_{k=0}^M P(D_{i,k} | A_{k,j^*}) \right) = S(c_i, T, \sigma_i) \end{aligned} \quad (16.4)$$

رابطه ۱۶.۴ همان‌طور که مشاهده می‌شود به راحتی قابل حل می‌باشد و نمونه‌ها مستقل از هم هستند و می‌توانند به درخت اتصال یابند اما ما تا به اینجا بهترین σ را به ازای یک j^* یافته‌ایم. آیا مقدار j^* نیز بهینه است؟ برای مشخص کردن مقدار بهینه j^* برای درخت دلخواه T از رابطه‌ای که در ادامه آمده است کمک می‌گیریم.

$$\langle \hat{\phi}, \hat{\sigma} \rangle = \arg \max_{\phi, \sigma} \prod_{i=1}^N S(c_i, T) \quad (17.4)$$

در واقع این مقادیر $\langle \hat{\phi}, \hat{\sigma} \rangle$ باید به گونه‌ای انتخاب شوند تا مجموع امتیازات همه اتصالات به درخت در حالت بیشینه خود باشد که برابر با امتیاز درخت می‌شود که در این حالت به $\hat{\phi}$ می‌رسیم که ماتریس A حاصل از آن را \hat{A} می‌نامیم. در نهایت که بهترین مقادیر به ازای درخت مشخص شدند پس می‌توان X را به صورت رابطه ۱۸.۴ تشکیل داد.

$$X_{i,j} = \hat{A}_{i,\hat{\sigma}_j} \quad (18.4)$$

۵.۴.۴ یافتن جهش‌های با پتانسیل حذف

همان‌گونه که از ابتدا می‌دانیم ما به دنبال درخت فیلوزنی حقیقی داده‌های نویزی مشاهده شده D هستیم. این درخت در این روش برابر با درختی است که،

- نحوه قرارگیری ژن‌ها در ساختار درخت (T)

- محل‌هایی در درخت که جهش‌های قبلی در آن‌ها حذف می‌شوند (ϕ)

- نحوه انتصاب نمونه‌های مشاهده شده به درخت (σ)

• و در نهایت پارامترهایی که برای مدل‌سازی خطای بوجود آمده در دادهای در دسترس مان تعیین شده است

(θ)

به گونه‌ای انتخاب شوند که محتمل‌ترین حالت را برای مشاهده داده‌ای D بوجود آورند که در این حالت ما رابطه علت توضیحات مجدد این موارد به این دلیل است که این بخش مهمترین بخش در ساختار روش پیشنهادی اول است.

۱.۵.۴.۴ مقایسه پاسخ فعلی با پاسخ آرمانی

پس از استخراج ماتریس مناسب از درخت می‌توان به ارزش‌گزاری و محاسبه درست‌نمایی پرداخت. این عمل به صورت رابطه ۱۹.۴ محاسبه می‌شود.

$$L : P(D|T, \sigma, \varphi, \theta) = \prod_{n=1}^{N} \prod_{m=1}^{M} P(D_{nm}|X_{nm}) \quad (19.4)$$

که X برابر ماتریس بدست آمده از درخت T با توجه به بردارهای σ و φ است. این رابطه بیانگر احتمال مشاهده ماتریس داده ورودی D در صورتی است که درخت فیلوزنی صحیح T و پارامترهای حقیقی θ باشد که توسط بردارهای σ و φ ثابت شده است. هرچه این احتمال بالاتر باشد نمایانگر این است که درخت، پارامترها و بردارهای کنترلی ما بگونه‌ای انتخاب شده‌اند که محتمل‌ترین حالت برای مشاهده داده‌های ورودی ما هست و در این صورت بهترین پاسخ برای ما همان پاسخی خواهد بود که محتمل‌ترین باشد. از این رو با دانستن θ ما به دنبال T ای به همراه بردارهای مربوطه آن هستیم که پاسخ رابطه باشد.

$$(T, \sigma, \varphi)_{\text{ML}} = \arg \max_{(T, \sigma, \varphi)} P(D|T, \sigma, \varphi, \theta) \quad (20.4)$$

اما همانگونه که می‌دانیم ما به دنبال بهترین درخت T هستیم که σ و φ در آن درخت برای ما اهمیت دارند. در واقع هر درخت T دارای امتیاز $S(T)$ است که به صورت رابطه تعریف می‌شود.

$$S(T) = P(D|T, \sigma^*, \varphi^*), \quad \sigma^* = \arg \max_{\sigma} P(D|T, \sigma, \varphi) \quad (21.4)$$

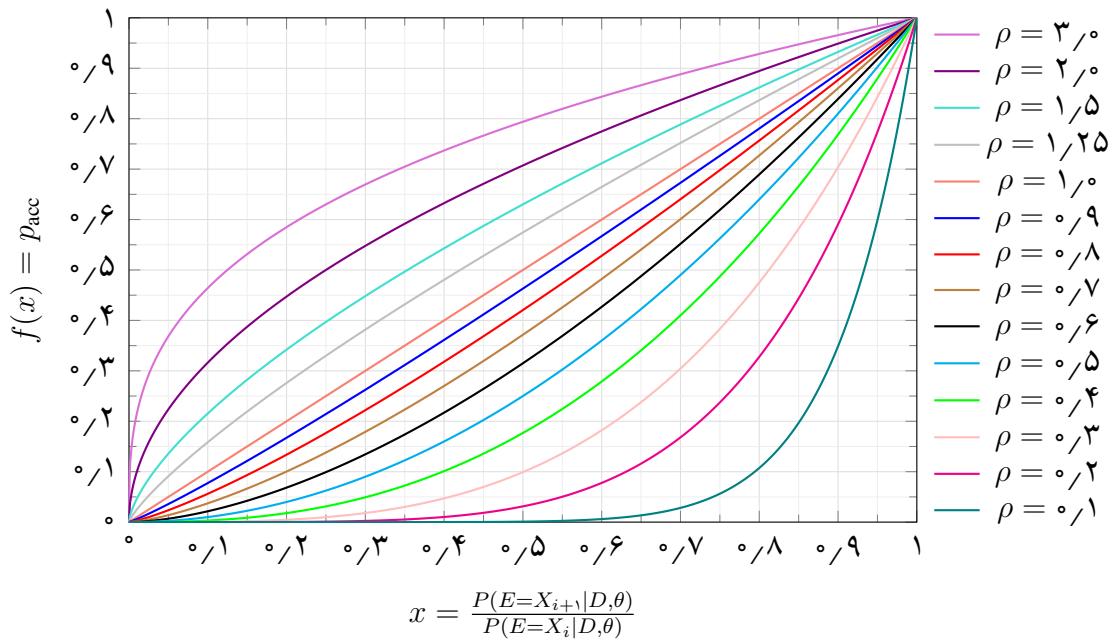
۴.۴.۶ پذیرش پاسخ‌های جدید و یافتن بهترین پاسخ

در این مرحله ما دو پاسخ با امتیازهایشان در اختیار داریم که می‌توانیم بر حسب آن‌ها برای ورود به تکرار بعد تصمیم‌گیری نماییم. این فرآیند توسط رابطه‌ای که در ادامه آمده است انجام می‌شود.

$$p_{\text{acc}} = \min \left[1, \left(\frac{P(E = X_{i+1} | D, \theta)}{P(E = X_i | D, \theta)} \right)^{\rho^{-1}} \right] \quad (22.4)$$

در رابطه ۲۲.۴، اگر پاسخ جدید بهتر از پاسخ فعلی باشد بیان می‌کند که افزایش بهینگی در پاسخ جدید باعث می‌شود تا صورت کسر مقداری بیش از مخرج بگیرید که در این صورت p_{acc} که برابر با احتمال پذیرش پاسخ جدید است، برابر ۱ خواهد شد که یعنی حتماً پاسخ جدید به عنوان پاسخ پابرجا برای ورود به تکرار بعد در نظر گرفته می‌شود. اما اگر پاسخ جدید (درخت جدید) بهتر از پاسخ فعلی ارزیابی نشود ما آن را مستقیماً رد نمی‌کنیم و به احتمالی کمتر از ۱ ممکن است آن را پذیریم. دلیل این پذیرش جلوگیری از به دام افتادن الگوریتم در پاسخ مان در بیشینه‌های محلی^۶ است. در شکل تاثیر تغییر پارامتر ρ در احتمال پذیرش پاسخ‌های جدیدی که مطلوب‌تر از پاسخ فعلی نیستند نمایش داده شده است.

⁶Local maxima


 شکل ۳.۴: نمودار تغییر احتمال پذیرش پاسخ جدید نامطلوب‌تر با توجه به مقدار پارامتر ρ .

۷.۴.۴ شبکه هرس‌کننده

در روش‌های گذشته مانند روش‌های ارائه شده در [۴۴] و [۸۲] در هر تکرار رویکرد MCMC از یک انتخاب کننده با توزیع یکنواخت برای انتخاب محل برش در درخت فعلی استفاده شده است. اما در روش پیشنهادی ارائه شده، سعی شده است تا این روش با یک روش هوشمند جایگزین شود که این عمل توسط یک شبکه عمیق یادگیری تقویتی انجام خواهد شد تا بتواند با انتخاب‌های هوشمند خود نسبت به حالت تصادفی فضای جستجو را کاهش داده و در نتیجه توانایی رسیدن به پاسخ مطلوب را با سرعت همگرایی بیشتر فراهم می‌کند. در ادامه این بخش به تشریح ساختار شبکه‌ای که برای مهم در نظر گرفته شده است پرداخته می‌شود.

۱.۷.۴.۴ ورودی

ورودی شبکه برابر ماتریس $\text{abs}(X - D)$ است که آن را I می‌نامیم. این ورودی که ابعادی برابر $N \times M$ دارد را به صورتی که در رابطه ۱.۷.۴.۴ آمده است به ماتریس جدید I' تبدیل می‌کنیم.

$$\begin{bmatrix} I_{1,1} & I_{1,2} & \dots & I_{1,j} & \dots & I_{1,M} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ I_{i,1} & I_{i,2} & \dots & I_{i,j} & \dots & I_{i,M} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ I_{N,1} & I_{N,2} & \dots & I_{N,j} & \dots & I_{N,M} \end{bmatrix}_{N \times M} \rightarrow \begin{bmatrix} 1 & 1 & f(1,1) & I_{1,1} \\ 1 & 2 & f(1,2) & I_{1,2} \\ \vdots & \vdots & \vdots & \vdots \\ i & j & f(i,j) & I_{i,j} \\ \vdots & \vdots & \vdots & \vdots \\ N & M & f(N,M) & I_{N,M} \end{bmatrix}_{N \times M \times 1}$$

در صورتی که

$$f(i,j) = \begin{cases} \alpha & \text{if } X_{i,j} - D_{i,j} = -1 \text{ (False Positive)} \\ 1 - \alpha & \text{if } X_{i,j} + D_{i,j} = 0 \text{ (True Negetive)} \\ \beta & \text{if } X_{i,j} - D_{i,j} = 1 \text{ (False Negetive)} \\ 1 - \beta & \text{if } X_{i,j} + D_{i,j} = 2 \text{ (True Positive)} \end{cases}. \quad (۲۳.۴)$$

همان‌گونه که مشخص است هر سطر از ماتریس I' دقیقاً برابر با یکی از درایه‌های ماتریس I است. دو ستون اول ماتریس I' متبار با اندیس‌های درایه‌های ماتریس I می‌باشد که ستون اول برابر با شماره سطر و ستون دوم برابر با شماره ستون است. ستون سوم ماتریس I' به احتمال وقوع و به نوعی تشریح‌کننده حقیقت ماجرا هست و حد خطا را مشخص می‌کند. به عبارت ساده‌تر مقادیر ممکن در ستون سوم چهار حالت مختلف می‌توانند داشته باشند که مقادیر α ، $1 - \alpha$ ، β و $1 - \beta$ هستند که بر حسب همچنین ستون آخر ماتریس I' برابر با مقدار درایه‌های ماتریس I است که دو مقدار 0 و 1 را خواهد داشت. بنابرین یا این روش توانستیم هر سطر از ماتریس جدید I' را متناظر با یک درایه از ماتریس ابتدایی درنظر بگیریم به‌طوری که حال تمام سطرها با یک دیگر متفاوت خواهند بود (بخاطر دو ستون اول که یکی به نمونه اشاره می‌کند و دیگری به جهش) در صورتی که در حالت اولیه در ماتریس

ابتداً یعنی درایه‌ها فقط دو یا نهایتاً چهار حالت مختلف می‌توانستند داشته باشند. این همان کلیدی هست که باعث می‌شود در ادامه شبکه قادر به یادگیری و دادن خروجی‌های مطلوب باشد.

۲.۷.۴.۴ ساختار شبکه

پس از مشخص شدن ورودی‌های شبکه نوبت به مشخص کردن ساختار شبکه و مأذول‌های استفاده شده در آن است. به همین منظور در ادامه این بخش به تشریح ساختار شبکه یادگیری تقویتی عمیق استفاده شده در این بخش می‌پردازیم.

در ابتدا یک بخش رمزگذار^۵ خواهیم داشت که ورودی‌ها را دریافت و آن‌ها را به بردار ویژگی^۶ تبدیل خواهد کرد. اگر ماتریس D ما به تعداد N نمونه (سطر) و M جهش (ستون) داشته باشد در آن صورت ماتریس I' ما دارای ابعاد $(M \times N, ۴)$ خواهد بود که در کل فضای حالت $4MN$ را خواهند داشت که MN بخاطر دو ستون اول ماتریس I' است که هر سطر با دیگری متفاوت است و 4 بخاطر دو ستون انتهایی ماتریس I' است که هر کدام 4 حالت مختلف را می‌توانند داشته باشند که در مجموع فضای حالت $4MN$ را می‌توانند اختیار کنند. هر داده از این ماتریس ورودی را به برداری به طول 2 embed می‌کنیم که خروجی برداری به ابعاد $(M \times N, ۲)$ خواهد شد که حال این اعداد شناور^۷ که در یک رنج وارزش هستند، بعد از یک تغییر بعد به صورت $(M \times N, ۸)$ حال این داده‌ها آماده استخراج ویژگی و اعمال عملیات کانولوشنی بر رویشان هستند. پس از تغییر بعد دادگان که

$(\text{Step 4}) \text{ have will we thus, convolutions, } 1D \text{ apply : }$

$\text{dim} * \text{hidden}_d \text{ and } 32 = 4 * \text{input} = \text{Notice:}$

$(\text{Step 5}) \text{ together representation two add : }$

resid- check: to structure inception like convolutions of connectivity the Check Idea:

pooling pyramid guided, attention ual.

⁷encoder

⁸Feature vector

⁹Float number

۸.۴.۴ شبکه بازاتصال‌کننده

۹.۴.۴ جمع‌بندی و نتیجه‌گیری

روش پیشنهادی ما الگو گرفته شده از روش ارائه شده در مقاله سایت بود اما با این تفاوت که در آنجا فرض مکان‌های بی نهایت بود ولی ما فرض اسکارلت را جایگزین کردیم که در این بین چون فضای جست و جو بزرگتر شد در نتیجه مجبور شدیم نحوه برداشتن گام‌های خود را تغییر بدیم و هوشمندانه تر جلو برویم که در این فضای بزرگتر بتوانیم به جواب مناسب برسیم. در سایت برای رسیدن به درخت بهتر به دنبال تنظیم کردن پارامترهای خطای بود در حالی که ممکن بود با جواب واقعی فاصله داشته باشد اما چون بدنبال جواب با امتیاز بالا بود در نتیجه این پارامتری بودن و جست و جو برای مقادیر بهینه خطای در روش آن وجود داشت اما ما برای بهتر کردن امتیاز به جای تغییر پارامترهای خطای ازای یک درخت جست و جوی ضمیمه کردن های مختلف سلول‌ها و لاس شدن جهش‌ها را جست و جو می‌کنیم.

در این مقاله الگوریتم **scarlet** معرفی شد که در آن به طور همزمان از دگرگونی تکهستهای (SNV) و جهش‌های حذف و تغییر تعداد کپی (CNA) از داده‌های توالی‌یابی تک سلولی برای استنباط فیلوزنی تومور استفاده شد. این الگوریتم، یک مدل تکاملی بر اساس در نظر گرفتن خطای ناشی از حذف جهش است که حذف جهش‌ها را محدود به مکان‌هایی می‌کند که شواهدی از حذف جهش‌های حذف و تغییر تعداد کپی موجود باشد. مدل‌های فیلوزنی با در نظر گرفتن حذف خطا، با استفاده از اطلاعات جهش‌های حذف و تغییر تعداد کپی که به آسانی در داده‌های دگرگونی تکهستهای موجود است، نسب به مدل‌های دولو یا فرض مکان‌های بی‌نهایت، ابهام کمتری در استنباط درخت فیلوزنی دارند. اگر چه به صورت طبیعی در داده‌های دگرگونی تکهستهای یک عدم قطعیت ذاتی در حضور یا عدم حضور جهش در سلول‌ها وجود دارد، اما کاهش میزان ابهام در استنباط فیلوزنی تومور منجر به افزایش دقت فیلوزنی استنباط شده است. در این مقاله نشان داده شد که فیلوزنی توموری استنباط شده برای بیماران مبتلا به سرطان روده از دقت و تکرارپذیری بیشتری برخوردار است و این الگوریتم در نهایت فیلوزنی‌هایی را استنباط کرد که در آن ۳ حذف جهش رخ داده بود. البته این الگوریتم محدودیت‌های خاص خود را دارد. به عنوان مثال، این نوع پیاده‌سازی از الگوریتم اسکارلت مستلزم درخت حذف و تغییر تعداد کپی به عنوان ورودی و میزان درست‌نمایی هر یک از این درختان است. این رویکرد در موقعي که تعداد مشخصی از تغییرات تعداد کپی وجود دارد قابل اجراست اما هنگامی که داده‌های توالی‌یابی تک سلولی در مقیاس بزرگ انجام شود، به درختان زیادی از جهش‌های حذف و تغییر تعداد کپی نیاز خواهد بود.

فصل ۵

نتایج تجربی

۱.۵ پایگاه داده‌های ورودی

قبل از اینکه وارد روش پیشنهادی شویم به تشریح وردی‌های مسئله و داده‌هایی که مورد استفاده قرار خواهیم داد می‌پردازیم. داده‌های ورودی برابر ماتریس $D_{m \times n}$ می‌باشد که بعد اول M برابر با ژن‌ها و بعد دوم N برابر سلول‌های نمونه‌برداری شده می‌باشد. در هر خانه $d_{i,j}$ یک بردار داده قرار دارد که حاوی اطلاعات ژن j در سلول i می‌باشد.

۱.۱.۵ پایگاه داده مصنوعی^۱

با توجه به این نکته که از درخت فیلوژنی حقیقی^۲ داده‌های حقیقی موجود اطلاعی نداریم، به سراغ ساخت پایگاه داده مصنوعی می‌رویم. با استفاده از این پایگاه داده مصنوعی می‌توانیم در مورد روش‌هایی که در ادامه بیان خواهیم کرد یک معیار ارزیابی نسبتاً مناسبی داشته باشیم و تا حدودی از مشکلات روش‌های پیشنهادی آکاه شویم و به تصحیح آن پردازیم. برای ساخت پایگاه داده مصنوعی که همان ماتریس ورودی $D_{m \times n}$ می‌باشد، از دو روش مختلف با دو فرض مختلف استفاده خواهیم کرد که در ادامه به تشریح هر کدام خواهیم پرداخت. برای ایجاد پایگاه داده در این حالت ابتدا درختی تصادفی با پارامترهای n ،^۳ ایجاد می‌کنیم که n تعداد ژن‌ها

¹Synthetic Dataset

²Ground-truth Phylogeny Tree

(جهش‌ها) بوده و عددی در بازه (۰، ۱۰۰) است که یک پارامتر کنترلی است که وظیفه اش کنترل کلی تعداد نسل‌های مختلف را از یک جمعیت در درخت فیلوزنی می‌باشد. حال برای تولید پایگاه داده مصنوعی به ترتیب سه گام زیر باید انجام شود.

- ایجاد یک درخت فیلوزنی تصادفی
 - تبدیل درخت فیلوزنی به ماتریس اطلاعات سلول-ژن (E)
 - اضافه کردن نویز به ماتریس E و تبدیل آن به ماتریس نویزی D
- در ادامه هر بخش به صورت جداگانه به تفضیل شرح داده خواهد شد.

۱.۱.۱.۵ ساخت درخت تصادفی

برای ساخت درخت تصادفی از دو روش مختلف استفاده شده است که هرکدام جداگانه توضیح داده شده است.

روش اول: با استفاده از درخت تصادفی دودویی ژنولوژی^۳

در این روش همان‌گونه که از نام آن مشخص است با استفاده از درخت تصادفی دودویی ژنولوژی به ساخت ماتریس داده ورودی مسله می‌پردازیم که برای ساخت این دادگان از فرض‌های که در ادامه آمده است استفاده خواهیم کرد.

در مرحله اول که ساخت درخت است به این صورت عمل می‌کنیم که به تعداد n گونه (سلول) در نظر می‌گیریم. سپس به ترتیب مراحل زیر را انجام می‌دهیم تا به درخت تصادفی مورد نظر برسیم.

- به هر کدام از n گونه متمایز در ابتدا وزن $w_i = 1$ را اختصاص می‌دهیم که متناسب با احتمال انتخاب هر گونه در مراحل بعدی خواهد بود.

• برای هر گونه i تابع جرم احتمال را در ادامه به صورت $F_i = \frac{w_i}{\sum_{i=1}^n w_i}$ در نظر می‌گیریم

- با استفاده از F دو گونه متمایز v, u را انتخاب می‌کنیم و به هم متصل می‌کنیم

³Random Binary Genealogical Tree

- به جای دو گونه u, v یک گونه جدید uv با وزن $w_{uv} = \frac{w_u + w_v}{\sqrt{2}}$ را قرار می‌دهیم.

• تعداد گونه‌ها یک واحد کم شده است. بررسی می‌کنیم اگر تعداد گونه‌های باقی‌مانده از ۲ کمتر باشد

درخت تصادفی ساخته شده است و پایان کار است. در غیر این صورت به مرحله اول بازمی‌گردیم.

پارامتر ζ به گونه‌ای کنترل‌کننده میزان ناپایداری در طی نسل‌ها می‌باشد. بطوریکه نمونه‌ای از نتایج مقادیر مختلف آن برای $n = 20$ در شکل ۱۰.۵ آورده شده است. پس از ساخت درخت تصادفی به سراغ مرحله بعد یعنی تبدیل درخت به ماتریس ژن-سلول E می‌رویم.

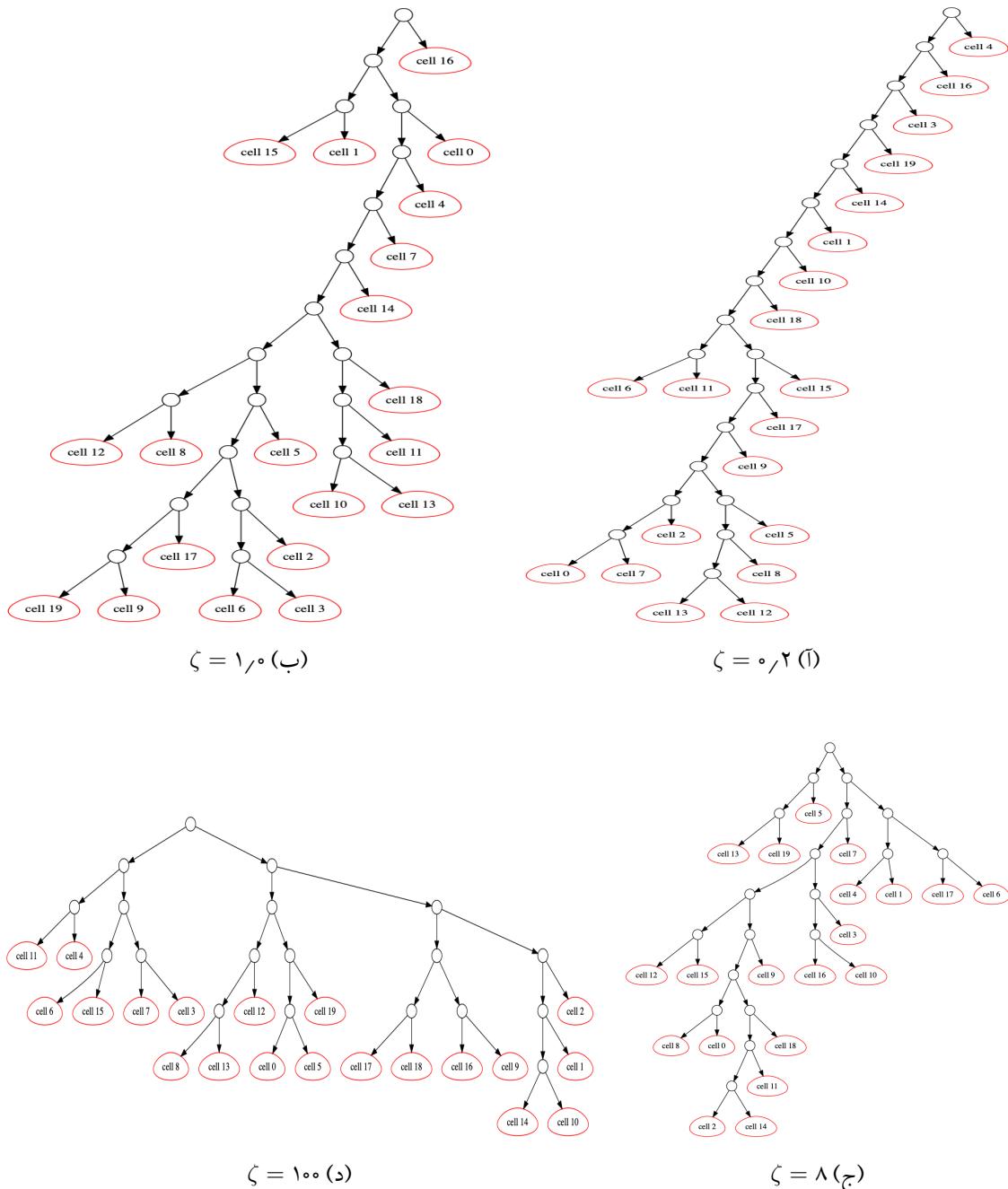
در ادامه با توجه به اینکه تعداد دلخواه جهش‌ها چه عددی بوده است یکی از گام‌های زیر را برمی‌داریم.

• اگر تعداد جهش‌ها $N > M$ بوده باشد در آن صورت به صورت تصادفی به تعداد دفعات اختلاف یکی از انشعاب‌ها در درخت را به صورت تصادفی انتخاب کرده و آن جهش اضافه شده را تا تمامی نوادگان پیش خواهیم برد.

• اگر تعداد جهش‌ها $N < M$ بوده باشد آنگاه مجدداً به اندازه تعداد اختلاف انشعاب‌هایی را انتخاب کرده و این بار جهش در آن انشعاب را تا تمامی نوادگان حذف می‌کنیم.

به این ترتیب تمامی سلول‌ها را با تعداد جهش‌های انتخابی خواهیم داشت. در نهایت برای اخیرین تغییر در جهش‌ها می‌توان یک گام دیگر برداشت که آن تولید یه عدد تصادفی کوچکتر از $\frac{M}{2}$ است که به آن تعداد می‌توان جهش‌های موجود را از انشعابی برداشت و بر روی انشعابی دیگر قرار داد. با این کار ممکن است تعداد جهش‌ها در انشعاب‌های مختلف تغییر کند و چه بسا به مدل‌های واقعی نزدیکتر شود که البته در این پایان‌نامه از گام آخر صرف نظر کرده‌ایم.

حال کار ما با پخش تصادفی جهش‌ها در پایگاه‌داده مجازی پایان یافته است. تا به اینجا ما در فرض خود از هر نمونه جمعیت مختلف یک سلول داشته‌ایم. اما در بعضی مواقع در پایگاه داده‌های واقعی ممکن است از یک جمعیت بیش از یک نمونه وجود داشته باشد که البته این امر لزوماً درست نیست به این دلیل که بعد از افزوده شدن نویز به داده‌ها ممکن است برخی سلول‌ها جهش‌هایشان مشابه هم شود. اما به هر حال اگر چنین چیزی را بخواهیم که داشته باشیم با انتخاب تصادفی برخی سلول‌ها (برگ‌ها) در درخت و کپی کردن آن‌ها می‌توان به چنین مقصودی رسید.



شکل ۱.۵: درخت فیلورژنی تصادفی تولید شده برای $n = 20$ و ζ ‌های مختلف

روش دوم: با استفاده از درخت تصادفی جهش‌های ژنی^۴

این روش نیز تا حدود زیادی مشابه روش قبل است با این تفاوت که در اینجا به جای اینکه درخت تصادفی را

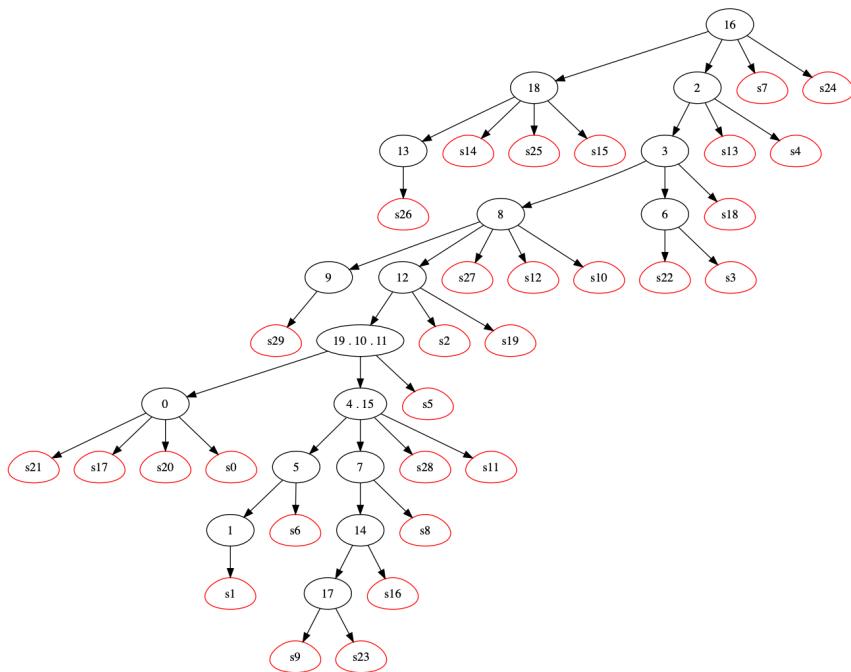
⁴Random Mutation History Tree

با توجه سلول‌ها از پایین به بالا بسازیم، ابتدا یک درخت تصادفی بدون در نظر گرفتن سلول‌ها ایجاد می‌کنیم و سپس به تخصیص جهش‌ها به آن می‌پردازیم و در نهایت برای آخرین مرحله به تعداد دلخواه سلول را به درخت اضافه کرده و درخت را تکمیل می‌کنیم. در گام اول به تعداد $1 + M$ نود در نظر می‌گیریم. مشابه حالت قبل با طی مراحلی بکه در ادامه آمده است به ساختار یک درخت تصادفی می‌رسیم.

- به هر کدام از m نود متمایز در ابتدا وزن $1 = w_i$ را اختصاص می‌دهیم که متناسب با روند حرکتی تومور به سمت آن جهش‌ها در مراحل بعدی خواهد بود.
- برای هر نود i تابع جرم احتمال را در ادامه به صورت $F_i = \frac{w_i}{\sum_{i=1}^n w_i}$ بیان می‌شود در نظر می‌گیریم.
- با استفاده از F دو نود متمایز v, u را انتخاب می‌کنیم و به هم متصل می‌کنیم.
- به جای دو گونه v, u یک نود جدید uv با وزن $\frac{w_u + w_v}{\sqrt{\zeta}} = w_{uv}$ را قرار می‌دهیم.
- تعداد نودها یک واحد کم شده است. بررسی می‌کنیم اگر تعداد نودهای باقیمانده از ۲ کمتر باشد به مرحله بعد می‌رویم و در غیر این صورت به مرحله اول بازمی‌گردیم.
- در این مرحله تمامی برگ‌های درخت ساخته شده را حذف می‌کنیم و تنها باقیمانده را به عنوان درخت تصادفی جهش‌ها در نظر می‌گیریم.

پس از به پایان رسیدن مراحلی که بیان شد درخت تصادفی آماده است و حال نوبت به تخصیص دادن خود زن‌ها به هر کدام از این نودهای درخت است. برای این منظور به هر کدام از M نود یک زن را به صورت تصادفی تخصیص می‌دهیم. پس از آن برای نهایی سازی درخت جهش‌ها از پارامتر دلخواه $\lfloor (1 - \gamma) * (M - 1) \rfloor$ استفاده می‌کنیم که γ عددی بین $(0, 1)$ است و A تعداد یال‌هایی است که در درخت باید برداشته شود و دو نود آن با یکدیگر ادغام شود. این کار باعث می‌شود تا در درخت جهش‌ها در برخی نودها به جای یک جهش چند جهش داشته باشیم که بتواند به مدل داده‌های واقعی نزدیکتر باشد.

پس از تکمیل درخت جهش‌ها نوبت قرار دادن نمونه‌هایی بر روی آن است. به همین منظور با فرض اینکه $N \geq M$ است. به تعداد M تا از سلول‌ها را به هر کدام از نودهای درخت جهش به عنوان برگ‌های جدید اضافه می‌کنیم و برای $N - m$ سلول باقیمانده همین کار را این‌بار به صورت تصادفی انجام می‌دهیم. در نهایت درخت تصادفی جهش‌ها ساخته شده است که نمونه‌ای از آن را در شکل ۲.۵ قابل مشاهده است.



شکل ۲.۵: درخت جهش تصادفی با پارامترهای $N = ۳۰, M = ۲۰, \zeta = ۱, \gamma = ۰, ۱۵$

۲.۱.۱.۵ تبدیل درخت به ماتریس ژن-سلول

با داشتن درخت (تولید شده با هر کدام از روش‌ها تفاوتی ندارد) در ادامه از فرض‌های مختلف در تولید ماتریس

E می‌توان استفاده کرد.

فرض مدل مکان‌های بی‌نهایت^۵

در این حالت فرض می‌کنیم که هر جهش اتفاق افتاده در درخت فیلوزنی در تمامی نسل‌های پس از آن باقی می‌ماند و هیچ‌گاه از بین نمی‌رود. در چنین حالتی درخت حاصل از این روش درختی یکتا بوده که به نام درخت فیلوزنی کامل^۶ شناخته می‌شود.

در این قسمت باید با استفاده از درخت تصادفی تولید بتوانیم ماتریس جهش‌ها را برای سلول‌های مختلف با فرض مکان‌های بی‌نهایت بدست آوریم. در ابتدا ماتریس E را به ابعاد $M \times N$ ایجاد می‌کنیم و برای هر درایه j, i در

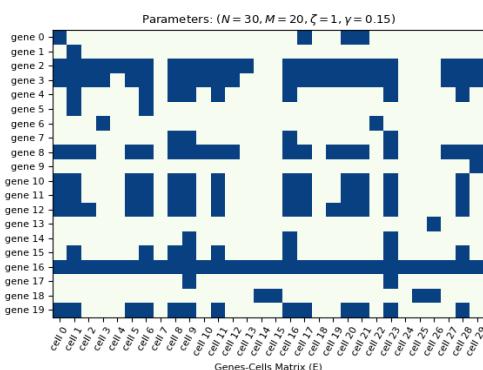
⁵Infinite Site Models

⁶Perfect Phylogeny Tree

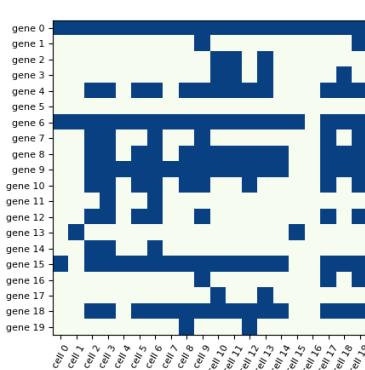
آن که \mathcal{N} شماره جهش و \mathcal{Z} شماره سلول است به صورت فرمولی که در ادامه آمده است مقداردهی می‌کنیم.

$$E_{i,j} = \begin{cases} 1 & \text{if mutation } i \text{ is an ancestor of cell } j \\ 0 & \text{o.w} \end{cases} \quad (1.5)$$

به این ترتیب با فرض مدل مکان‌های بینهایت ماتریس بدون خط E را داریم که برای تصاویر دوروش درخت مرحله قبل در شکل ۳.۵ بدست آمده‌اند.



(ا) ماتریس درخت شکل ۱.۵



(ب) ماتریس درخت شکل ۲.۵

شکل ۳.۵: ماتریس‌های ژن-سلول (E) بدست آمده از درخت‌های تصادفی ساخته شده

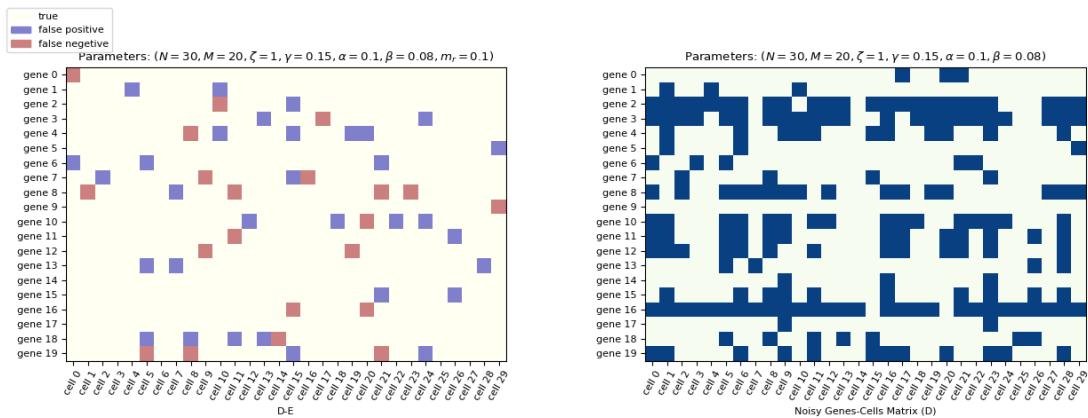
۳.۱.۱.۵ اضافه کردن نویز به ماتریس ژن-جهش

برای قسمت نهایی آمده سازی پایگاه داده مجازی نیاز است تا به ماتریس E با پارامتر $\Theta = (\alpha, \beta, m_r)$ نویز اضافه کنیم و آن را به ماتریس D تبدیل کنیم که $\beta = P(D_{ij} | E_{ij} = 1)$ و $\alpha = P(D_{ij} = 1 | E_{ij} = 0)$ است و همچنین $m_r \in (0, 1)$ که نرخ داده‌های از دست رفته را مشخص می‌کند.

برای این منظور به ازای تمامی درایه‌های 0 ماتریس E هر بار یک عدد تصادفی با توزیع یکنواخت بین $(0, 1]$ بوجود می‌آوریم و اگر عدد تولید شده کوچکتر از α بود آنگاه ان درایه در ماتریس D را برابر با 1 قرار می‌دهیم. به همین ترتیب مجدداً این بار برای درایه‌های 1 ماتریس E این کار را تکرار می‌کنیم و اگر عدد تصادفی تولید شده کوچکتر از β شد، درایه متناظر را در ماتریس D برابر با 0 قرار می‌دهیم.

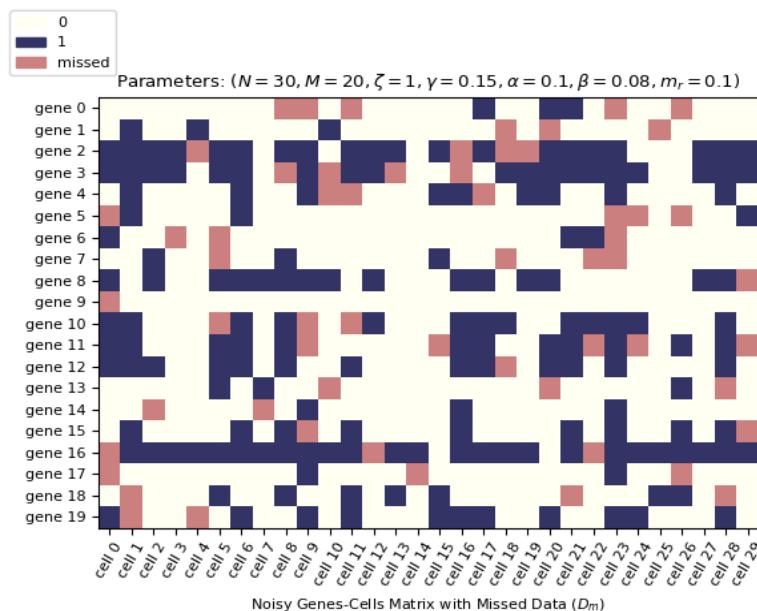
پس از اتمام کار نوبت به اضافه کردن داده‌های از دست رفته است. برای این منظور با نرخ m_r بعضی از درایه‌های ماتریس D را برابر با ۲ قرار می‌دهیم که به منزله در دسترس نبودن اطلاعات است. نام ماتریس نهایی را که شامل داده‌های از دست رفته است D_m می‌گزاریم. در ادامه تصاویر اضافه شدن نویز به ماتریس شکل ۳.۵ ب در شکل ۴.۵ آمده است.

۴.۵ آمده است.



(ب) نویزی اضافه شده با پارامترهای $\alpha = ۰/۱, \beta = ۰/۰۸$

(آ) ماتریس نویزی با $\alpha = ۰/۱, \beta = ۰/۰۸$

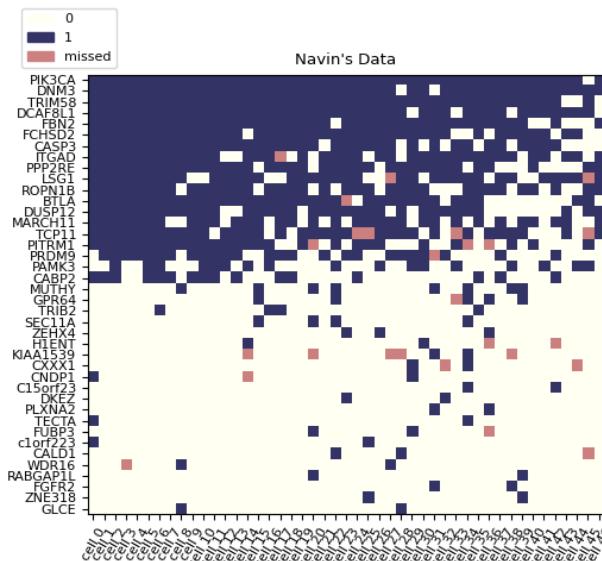


(ج) ماتریس نویزی به همراه داده‌های از دست رفته با پارامترهای $\alpha = ۰/۱, \beta = ۰/۰۸, m_r = ۰/۱$

شکل ۴.۵: ماتریس‌های ژن-سلول همراه با نویز و داده‌های از دست رفته شکل ۳.۵ ب که برای ورودی مسله آماده شده است.

۲.۱.۵ پایگاه داده حقیقی^۵

به عنوان پایگاه داده حقیقی از پایگاه داده استفاده شده در مقاله SCITE به عنوان پایگاه داده حقیقی اصلی استفاده خواهیم کرد که ماتریس داده ورودی آن به صورت شکل ۵.۵ می‌باشد. همچنین پایگاه داده حقیقی ^۶Xu



شکل ۵: داده‌های حقیقی Navin در مقاله SCITE

نیز که در مقاله SCITE مورد استفاده قرار گرفته است در شکل ۶.۵ آمده است.

⁷Real Dataset



شکل ۶.۵: داده‌های حقیقی Xu در مقاله SCITE

۲.۵ روش پیشنهادی بدست آوردن درخت فیلوزنی

پس از تخمین داده‌های از دست رفته، در این بخش به معرفی روش پیشنهادی برای یافتن درخت فیلوزنی می‌پردازیم. در روش‌های گذشته که رویکرد آن در ادامه بیان شده است به موفقیت نرسیدیم و این‌بار در نظر داریم تا با استفاده از یک درخت در ساختار شبکه ژن‌ها بتوانیم به یک درخت فیلوزنی مناسب دست یابیم.

- استفاده از شبکه ژن‌ها

۷ استفاده از یک گراف ابتدایی و سپس تغییر و هرس کردنش تا رسیدن به درخت جهش‌ها

۳ استفاده از یک درخت نمونه نابهینه و تغییر اتصالات تا رسیدن به درخت بهینه جهش‌ها

- استفاده از شبکه سلول‌ها

۷ استفاده از یک گراف سلول‌ها و بهینه‌کردن ارتباطات بین آن‌ها و سپس تبدیل آن به درخت فیلوزنی

در رویکرد اول ما از شبکه‌های ژنی استفاده خواهیم نمود. این شبکه‌ها نودهایی معادل با یک ژن متمایز را در نظر می‌گیرند. در گذشته با استفاده از شبکه‌ای کامل با وزن‌های متفاوت که بر حسب اطلاعات ورودی به الگوریتم تعیین می‌شد، متساقانه به موفقیت خاصی نرسیدیم. همچنین مشابه همین رویکرد را در ساختار

شبکه‌های سلولی دنبال کردیم که مجدداً پیشرفت قابل ملاحظه‌ای حاصل نشد. به همین جهت این‌بار در این گزارش با تغییری اساسی به دنبال یافتن روشی مناسب برای استنتاج درخت فیلوزنی می‌باشیم.

۱.۲.۵ استفاده از شبکه ژن‌ها برای یافتن درخت فیلوزنی

در این رویکرد با استفاده از شبکه‌ای که نودهایی معادل ژن‌ها داشته باشد سعی داریم تا به درخت فیلوزنی

بهینه برسیم.

۱.۱.۲.۵ استفاده از یک درخت نمونه نابهینه و تغییر اتصالات تا رسیدن به درخت فیلوزنی بهینه

در این روش قصد داریم تا با شروع از یک درخت نمونه که در ابتدا به صورت تصادفی از اتصال ژن‌ها بوجود آمده است، به بهینه‌ترین درخت ممکن برسیم. این روش به صورت تکرارواره با تغییر اتصالات درخت سعی در بدست آوردن درختی مطلوب‌تر دارد که شرایط و روابط تاثیرگزار در آن به تفضیل شرح داده خواهد شد. در واقع این روش پیشنهادی یک جستجوی حریصانه می‌باشد که طی شرایطی می‌توان انتظار داشت که به پاسخ بهینه دست یافته شود. این روش به نام روش زنجیره مارکو مونت-کارلو^۸ شناخته می‌شود که در بسیاری از مقالات مرتبط نیز مورد استفاده قرار گرفته شده است.

برای شروع یک درخت تصادفی T را با نودهایی معادل ژن‌های پایگاه داده ورودی در نظر می‌گیریم که در گام اول به صورت تصادفی ساخته شده است. در گام‌های بعدی یک نود n_1 را از درخت T به صورت تصادفی انتخاب می‌کنیم. سپس زیردرخت با ریشه این نود را از درخت کم می‌کنیم. حال در درخت باقی‌مانده یک نود دیگر n_2 را به صورت تصادفی انتخاب می‌کنیم و آن زیر درخت قبلی با ریشه n_1 را به n_2 متصل می‌کنیم و درخت جدید را T_n نام‌گذاری می‌کنیم. پس از آن با احتمال،

$$P = \min \left(1, \frac{Eng(T)}{Eng(T_n)} \right) \quad (2.5)$$

درخت جدید بدست آمده T_n را به عنوان نتیجه این گام می‌پذیریم و در غیر این صورت درخت این گام نیز همان درخت سابق T باقی خواهد ماند. در رابطه ۲.۵،تابع Eng برای یک درخت در واقع انرژی آن درخت را محاسبه

⁸Markov Chain Monte Carlo

می‌کند و ما به دنبال پایدارترین درخت هستیم که کمترین انرژی را داشته باشد. تعریف اینتابع برای یک درخت به این صورت است که با توجه به نمونه‌هایی که در دادگان ورودی D قرار دارد و اینکه کدام ژن بالاتر یا پایین‌تر از دیگر ژن‌ها قرار دارد به درخت یک نمره انرژی منصوب می‌کند که به صورت فرمول ۳.۵ بیان می‌شود.

$$Eng(T) = ||E - \hat{E}|| \quad (3.5)$$

که در اینجا \hat{E} ماتریس تخمین زده شده روش پیشنهادی با توجه به درخت نهایی بدست آمده خواهد بود. در واقع ماتریس E همان ماتریس صحیح بدون خطای مختلف است که جهش‌های مختلف را به ازای سلول‌های مختلف مشخص می‌کند. هنگامی که ما درخت ساخته شده فرضی T را داشته باشیم می‌توانیم در دو گام به \hat{E} برسیم. توجه به این نکته ضروری است که اگر در واقعیت فرض ما که همان مکان بی‌نهایت بود کامل برقار باشد و E را داشته باشیم، حتماً باید بتوانیم به درختی با $= 0$ $Eng(T)$ دست یابیم. اما از آنجایی که ما D را به عنوانی از تخمین E داریم بنابرین محاسبه خطای واقعی خواهد بود نه خود آن که در اصل به صورت فرمول ۴.۵ می‌شود.

$$Eng(T) \approx Err(T) = ||D - \hat{E}|| \quad (4.5)$$

می‌دانیم که هر نود از این درخت T یک مکان برای اتصال سلولی می‌تواند باشد که در این صورت معنی آن اینگونه خواهد بود که سلول ضمیمه شده به آن نود تمام جهش‌های والد خود را داشته است. بنابرین در گام اول نیاز است تا هر مکان از درخت مشخص شود که چه نمونه‌هایی می‌تواند تولید نماید. این اطلاع توسط ماتریس A مشخص می‌شود که به صورت زیر از روی درخت ساخته خواهد شد.

$$A_{i,j} = \begin{cases} 1 & \text{اگر } j = i \text{ یا جهش } i \text{ والد جهش } j \text{ باشد} \\ 0 & \text{در غیر این صورت} \end{cases} \quad (5.5)$$

حال در گام دوم کافی می‌توانیم با توجه به یک معیار بهترین انتخاب را برای ضمیمه کردن سلول‌ها (نمونه‌های) موجود به درخت داشته باشیم. به همین جهت در ماتریس D که هر ستون آن برابر با نمایش یک سلول است، می‌تواند با هر ستون از ماتریس A مقایسه شود و بهترین ستونی که از A انتخاب شود برابر با جایگاه مناسب

ضمیمه شدن نمونه با مقداری خطاب درخت T است. حال با توجه به اینکه فرض مکان‌های بی‌نهایت را داشتیم ماتریس E را به صورت زیر می‌سازیم.

$$\hat{E}_{i,j} = A_{i,\sigma_j} \quad (6.5)$$

که σ_i برابر با بهترین نود (زن) برای اتصال نمونه بردار d_j است که بهترین جایگاه به صورت فرمول زیر انتخاب می‌شود.

$$\begin{aligned} \sigma_j = \arg \max_{x \in [1 \rightarrow M]} & \sum_{i=1}^M \left[\right. \\ & A_{i,x} D_{i,j} (1 - \beta) + (1 - A_{i,x}) (1 - D_{i,j}) (1 - \alpha) + \\ & \left. A_{i,x} (1 - D_{i,j}) \beta + (1 - A_{i,x}) D_{i,j} \alpha \right] \end{aligned} \quad (7.5)$$

حال با داشتن ماتریس \hat{E} می‌توان خطای درخت را محاسبه نمود و با هدایت mcmc طبق فرمول ۸.۵ به درخت بهینه T_{op} رسید.

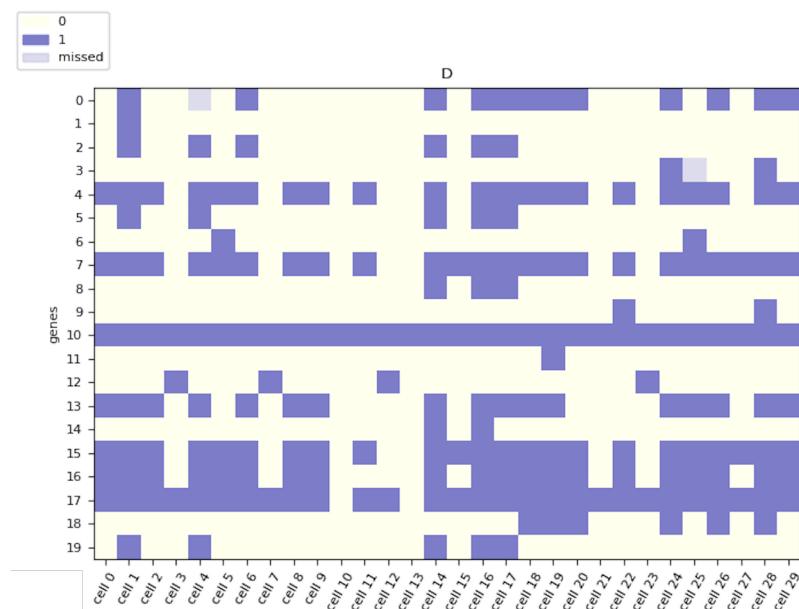
$$T_{op} = \min_{T \in \text{All possible } T} \left(\|D - \hat{E}_T\| \right) \quad (8.5)$$

۳.۵ نتایج تجربی

در این بخش به نتایج بدست آمده برای روش پیشنهادی می‌پردازیم و برای هر دو داده مصنوعی و حقیقی نتایج بدست آمده را تحلیل خواهیم نمود.

۱.۳.۵ نتایج بر روی پایگاه داده مصنوعی

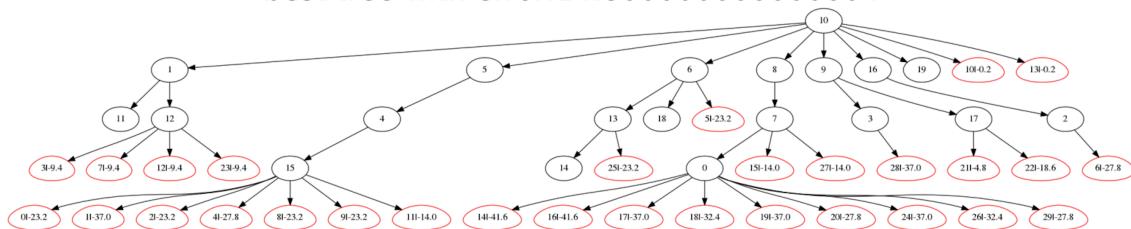
همان‌گونه که در بخش دوم توضیح داده شد با توجه به سختی دسترسی به پایگاه داده‌های حقیقی و اینکه در آن‌ها نیز حقیقت داده‌ها (E) وجود ندارد تصمیم به ایجاد پایگاه داده‌ای مصنوعی گرفته شد که با کمک آن بتوان ارزیابی مناسبی از روش پیشنهادی و میزان کارایی و مقاومت روش را نسبت به تغییر پارامترها سنجید. فرض کنید ماتریس ورودی شکل ۷.۵ را در اختیار داریم و میخواهیم بهترین درخت فیلوزنی را برای آن بیابیم.



شکل ۷.۵: نمونه‌ای تصادفی از ماتریس ورودی D

حال یک درخت تصادفی به صورت شکل ۸.۵ می‌سازیم. در درخت شکل ۸.۵ نمونه‌ها (سلول‌ها) با رنگ قرمز به درخت متصل شده‌اند که البته این ضمیمه بهترین ضمیمه ممکن است و میزان انرژی (خطای) هر ضمیمه نیز در کادر قرمز رنگ سلول‌ها به صورتی عددی منفی نوشته شده است.

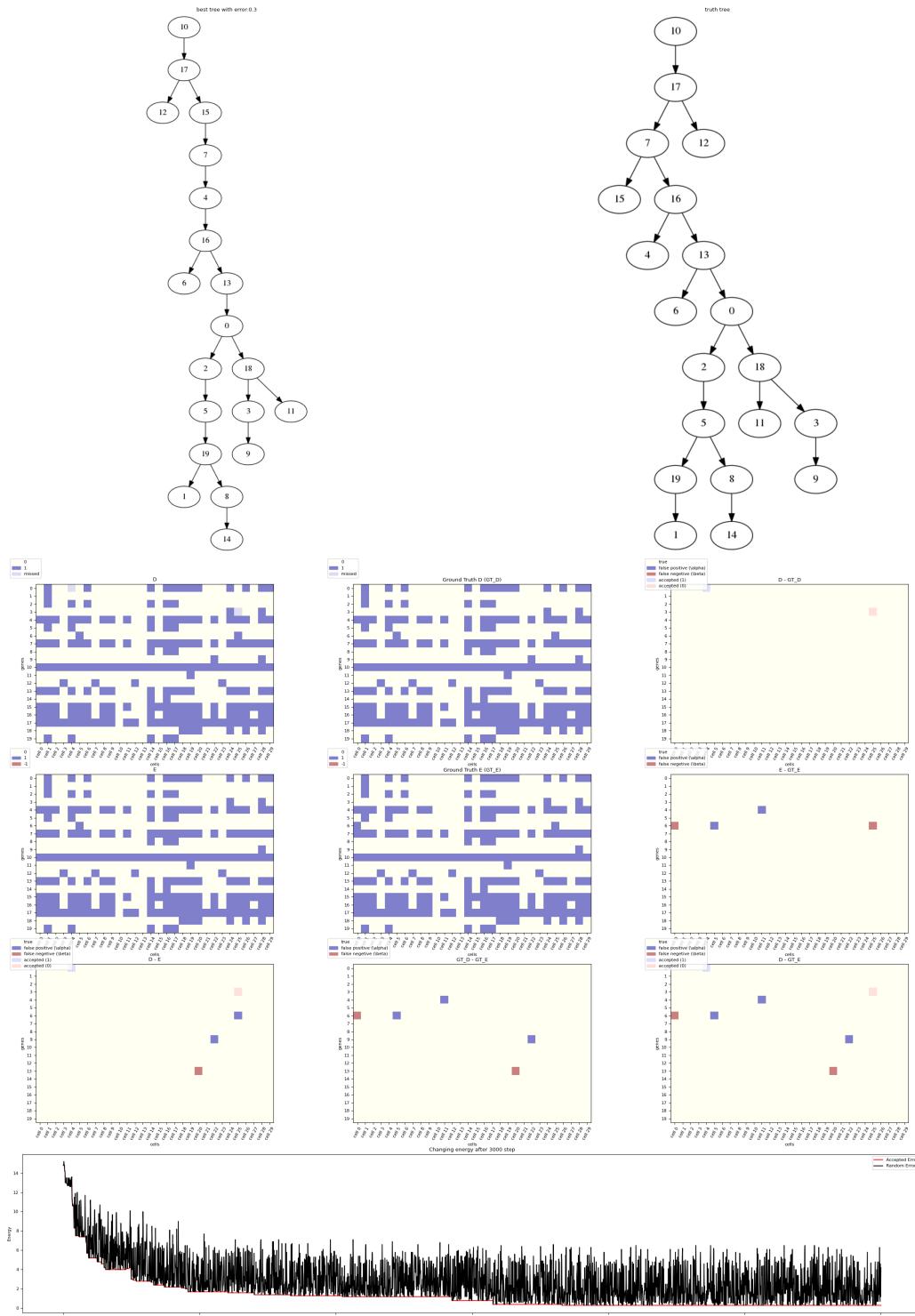
best tree with error:14.800000000000004



شکل ۷.۵: درخت تصادفی ایجاد شده به عنوان درخت اولیه شکل ۷.۵

پس از این مرحله اگر ۳۰۰۰ گام MCMC را اجرا نماییم می‌توانیم نتیجه حاصله را در شکل ۹.۵ مشاهده کنیم. در این شکل دو درخت وجود دارد که درخت سمت راستی درخت حقیقی است که به دنبال آن بودیم و درخت سمت چپ بهترین درخت یافته شده است. همچنین در پایین شکل، ۹ ماتریس مشاهده می‌شود که ماتریس‌ها سمت راست و پایین به نوعی بیان‌کننده میزان خطای بین ۴ ماتریس سمت چپ بالا می‌باشند. در بالای هر ماتریس نام آن نوشته شده است و در نهایت در انتهای تصویر نیز روند کاهش خطای تلاش‌های MCMC در گام‌های مختلف قابل مشاهده است. فقط نکته‌ای که وجود دارد این است که خطای نوشته شده در تصاویر برابر $1/\sqrt{0}$ مقیاس نوشته شده است.

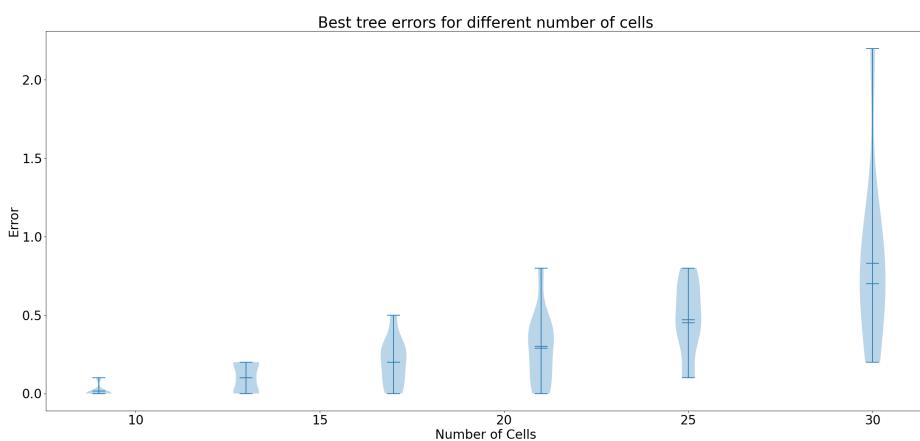
همان‌گونه که مشخص است در ماتریس D دو داده از دست رفته وجود دارد که یکی از آن‌ها در حقیقت جهش یافته و دیگر خیر. اگر ما در محاسبات خود این دو داده را در محاسبه خطای نظر نگیریم و با تغییر ۳ داده دیگر می‌توانیم به ماتریس \hat{E} (در شکل به نام E نوشته شده است) بررسیم که معادل بهترین درخت بدست آمده است. که این یعنی ماتریس D ما با ۵ تغییر بدست ما رسیده است. حال اگر حقیقت داده‌ها و درخت اصلی را مشاهده کنیم می‌بینیم که در آنجا نیز ۵ خطای وارد شده است که ۲ تای آن‌ها را درست کشف شده است. بنابرین الگوریتم بدون اطلاع از حقیقت توانسته با حداقل ۵ خطای یک درخت فیلوزنی مناسب دست بیابد که در ساختار نیز شباهات بسیار زیادی به حقیقت دارد. بنابرین روش پیشنهادی توانسته درخت فیلوزنی را با صحت $9916\% = \frac{2030-5}{2030}$ بازسازی کند که عددی قابل قبول می‌باشد.



شکل ۹.۵: نتیجه اجرای روش پیشنهادی برای ماتریس شکل ۷.۵

اما برای بررسی مناسب‌تر تعدادی تست را به ازای M و N ‌های مختلف اجرا نمودیم که به صورت خلاصه نتایج حاصل از آن در ادامه قابل مشاهده است.

در شکل ۱۰.۵ مقدار خطای درخت بهینه یافته شده قابل مشاهده است که نشان می‌دهد هر چه تعداد نمونه‌ها افزایش پیدا می‌کند و اندازه ماتریس ورودی بزرگ‌تر می‌شود، مقدار خطای نیز افزایش می‌یابد. در این اجرا تعداد جهش‌ها نیز عددی بین تعداد نمونه‌ها و نصف تعداد نمونه‌ها بوده است. حال برای اینکه متوجه شویم آیا این

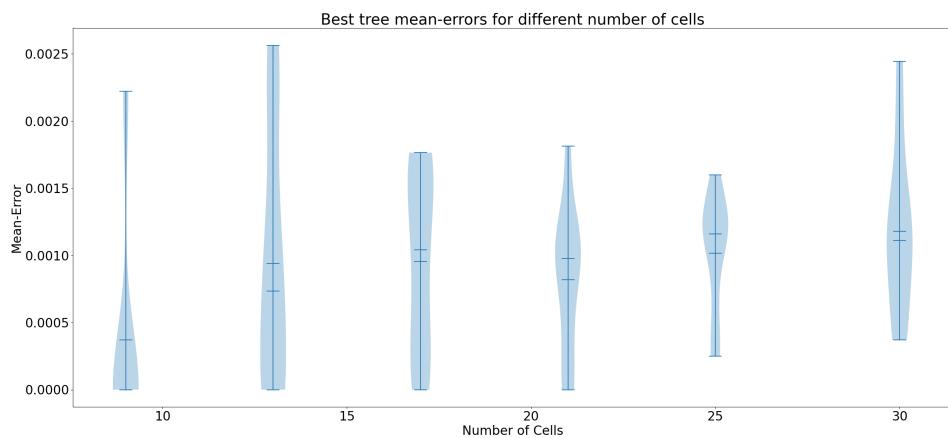


شکل ۱۰.۵: نتیجه اجرای روش پیشنهادی برای تعداد نمونه‌های مختلف

افزایش خطاب خاطر ضعف روش پیشنهادی است یا ماهیت داده‌های ورودی میزان خطای در هر اجرا بر تعداد خانه‌های ماتریس D تقسیم می‌کنیم که در آن صورت به نمودار شکل ۱۱.۵ می‌رسیم. در این نمودار جدید مشخص می‌شود که با افزایش اندازه ماتریس ورودی روش پیشنهادی سعی می‌کند تا خطای را به ازای هر داده کنترل کند که نشان از کارآمدی روش پیشنهادی می‌باشد.

۲.۳.۵ نتایج بر روی داده‌های حقیقی

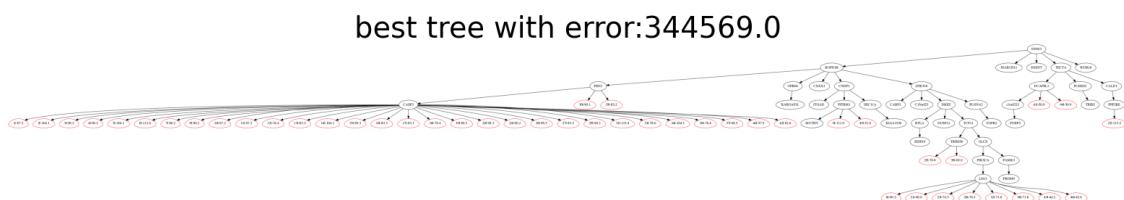
در این قسمت به ارائه گزارش و نتایج حاصل از روش‌های پیشنهادی با استفاده از داده‌های حقیقی برای بدست آوردن درخت فیلوزنی خواهیم پرداخت.



شکل ۱۱.۵: نتیجه اجرای روش پیشنهادی برای تعداد نمونه‌های مختلف

۱۰.۳.۵ نتایج بهینه‌سازی درخت ثُنی

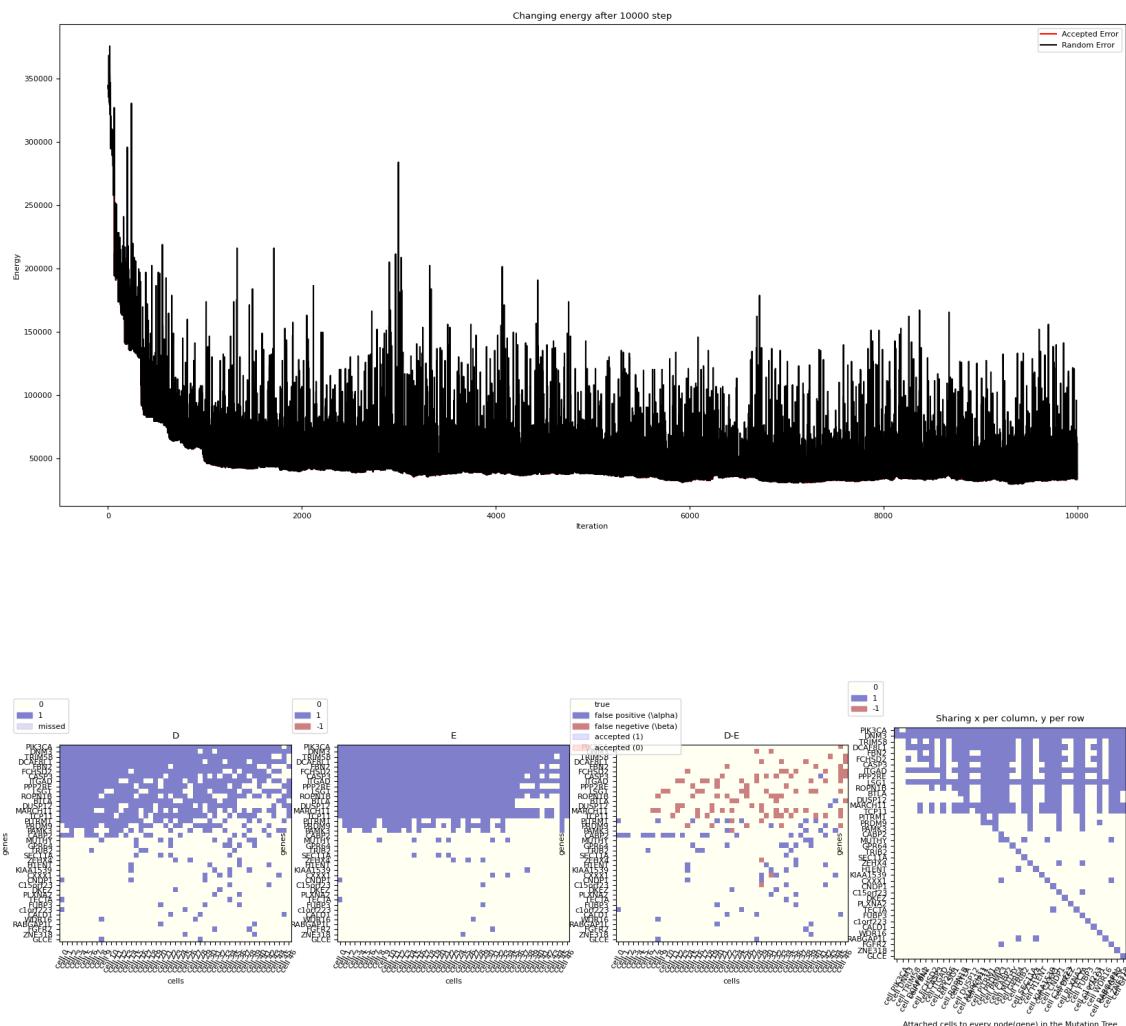
همان‌طور که در فصل قبل بیان شد، یکی از روش‌های بدست آوردن درخت فیلورژنی استفاده از یک درخت تصادفی نابهینه ژنی بود که طی تکرار گام‌هایی سعی در تغییر اتصالات و یافتن درخت بهینه داشت که بتواند روند صحیح تغییرات ژنی را در تومور مورد نظر نمایش دهد. نتیجه بدست آمده بر روی پایگاه داده حقیقی Navin به شرح زیر می‌باشد که عکس ۱۲.۵ درخت تصادفی اولیه الگوریتم را نشان می‌دهد که انرژی آن نیز بالای تصویر نوشته شده است.



شکل ۱۲.۵: درخت تصادفی اولیه

تصویر ۱۳.۵ نیز نمودار تغییر انرژی را طی گام‌های مختلف نمایش پیشنهادی مشخص می‌کند. در نهایت تصویر بهترین درخت یافته شده به همراه انرژی آن.

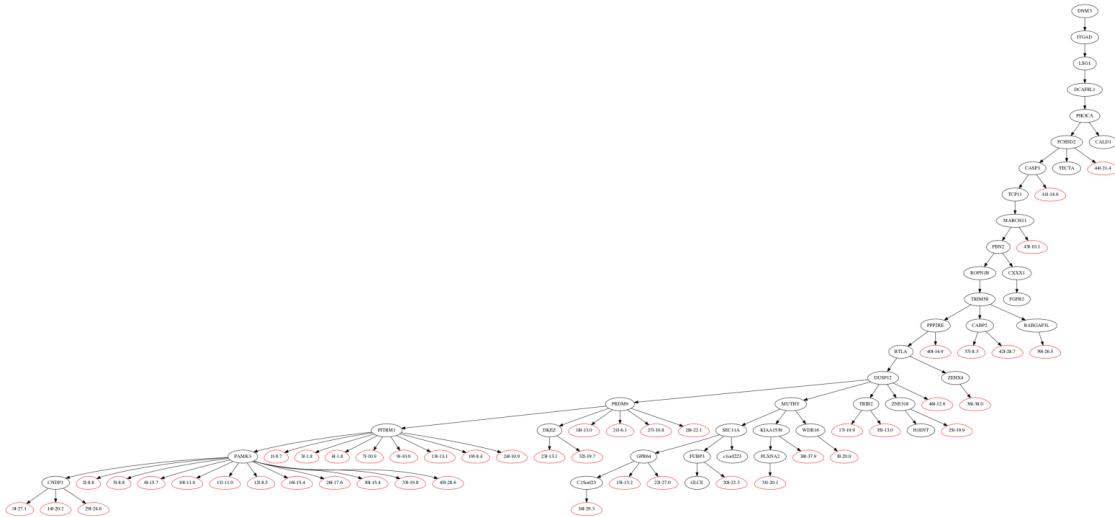
در نهایت برای مقایسه نیز تصویر درخت حاصله در مقاله اصلی SCITE را در شکل ۱۵.۵ نمایش داده شده



شکل ۱۳.۵: نمودار تغییر انرژی در طی گام‌های مختلف

است. همان‌طور که مشخص است مقدار انرژی بدست آمده برای خروجی الگوریتم پیشنهادی بهتر (کمتر) از انرژی درخت SCITE می‌باشد که دلیل بر بهینه‌تر بودن درخت روش پیشنهادی ارائه شده در این گزارش است.

best tree with error:29929.0



شکل ۱۴.۵: بهترین درخت یافته شده و خروجی الگوریتم برای مقاله SCITE

۴.۵ گام‌های آتی

در ادامه برای تکمیل روش پیشنهادی در دو قسمت نیاز به بهبود وجود دارد.

قسمت اول مربوط به درخت اولیه است و قسمت دیگر مربوط به سرعت MCMC می‌باشد.

۱.۴.۵ بهبود در ساخت درخت اولیه

در حال حاضر ما درخت اولیه را به صورت تصادفی انتخاب می‌کنیم که می‌توان در این مرحله درخت اولیه را با استفاده از مفروضات مدل مکان‌های بینهایت و با توجه به ماتریس ورودی بهبود بخشید. این کار باعث می‌شود تا شروع الگوریتم از نقطه بهتری باشد که در این صورت هم گام‌های لازم برای رسیدن به درخت بهینه می‌تواند کمتر شود و هم اینکه احتمال قرار گرفتن در نقاط اکسترم نسبی را کاهش می‌دهیم.

۲.۴.۵ افزایش سرعت همگرایی MCMC

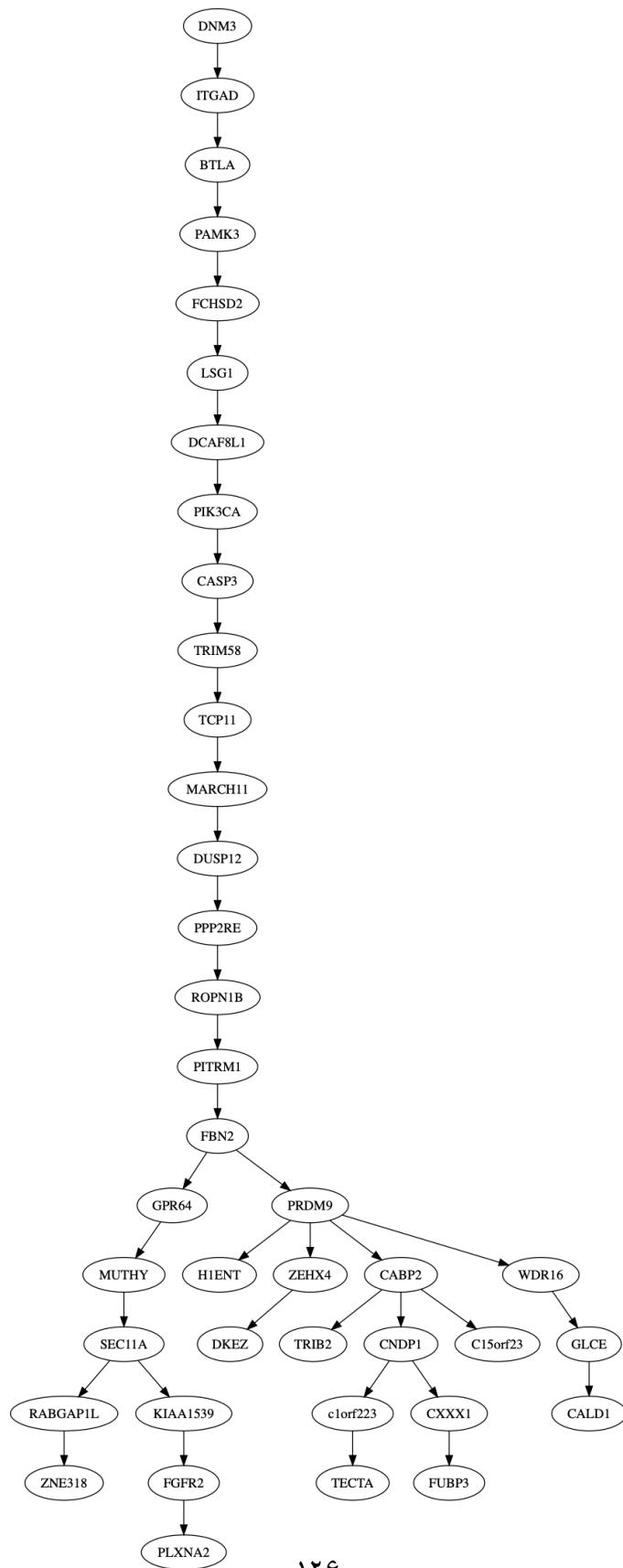
در این بخش نیاز است تا در دو قسمت روش پیشنهادی بهبود یابد.

۱.۲.۴.۵ توع در گام‌ها با استراتژی معقول

برای این بخش کاری که باید انجام شود این است که بتوان برای افزایش سرعت همگرایی از روش‌های مختلف در گام‌ها استفاده کرد. برای مثال در حال حاضر می‌توان از سه روش مختلف در هر گام استفاده نمود. روش اول تعویض دو نود در درخت می‌باشد. روش دوم جدایی یک زیر درخت و اتصال آن به محلی دیگر می‌باشد و در نهایت روش سوم تعویض دو زیر درخت با یکدیگر می‌باشد. با انتخاب یک استراتژی مناسب بین هرکدام از این روش‌ها در گام‌های مختلف احتمالاً بتوان سرعت همگرایی را افزایش داد.

۲.۲.۴.۵ قرار دادن احتمال وزن‌دار به ازای هر انتخاب

در حال حاضر ما در هرکدام از روش‌های مختلف که در بخش قبل برای گام‌های MCMC بیان کردیم، انتخاب نودها را به صورت کاملاً یکنواخت انجام می‌دهیم. در صورتی که احتمالاً بتوان با تعریف فرمولی مناسب این احتمال انتخاب بین نودهای مختلف در درخت را از حالت یکنواخت خارج کرد و در نتیجه مجدداً سرعت همگرایی الگوریتم را افزایش داد.



فصل ٦

بحث و نتیجه‌گیری

مراجع

- [1] Nci dictionary of cancer terms: somatic mutation definition. <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/somatic-mutation?redirect=true>.
- [2] Ii neoplasms. 19 June 2014.
- [3] Cancer - activity 1 - glossary. page page 4 of 5, 2008.
- [4] Abrams, Gerald. Neoplasia i. 23 January 2012.
- [5] Akselrod-Ballin, Ayelet, Karlinsky, Leonid, Hazan, Alon, Bakalo, Ran, Horesh, Ami Ben, Shoshan, Yoel, and Barkan, Ella. Deep learning for automatic detection of abnormal findings in breast mammography. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 321–329. Springer, 2017.
- [6] Alberts, Bruce, Johnson, Alexander, Lewis, Julian, Raff, Martin, Roberts, Keith, and Walter, Peter. Molecular biology of the cell 4th edition. New York: Garland Science, 1463, 2002.
- [7] Anderson, Kristina, Lutz, Christoph, Van Delft, Frederik W, Bateman, Caroline M, Guo, Yanping, Colman, Susan M, Kempski, Helena, Moorman, Anthony V, Titley, Ian, Swansbury, John, et al. Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature*, 469(7330):356–361, 2011.
- [8] Andor, Noemi, Graham, Trevor A, Jansen, Marnix, Xia, Li C, Aktipis, C Athena, Petritsch, Claudia, Ji, Hanlee P, and Maley, Carlo C. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature medicine*, 22(1):105–113, 2016.
- [9] Azer, Erfan Sadeqi, Ebrahimabadi, Mohammad Haghiri, Malikić, Salem, Khardon, Roni, and Sahinalp, S Cenk. Tumor phylogeny topology inference via deep learning. *Iscience*, 23(11):101655, 2020.

- [10] Beerenwinkel, Niko, Schwarz, Roland F, Gerstung, Moritz, and Markowetz, Florian. Cancer evolution: mathematical models and computational inference. *Systematic biology*, 64(1):e1–e25, 2015.
- [11] Behjati, Sam, Huch, Meritxell, van Boxtel, Ruben, Karthaus, Wouter, Wedge, David C, Tamuri, Asif U, Martincorena, Iñigo, Petljak, Mia, Alexandrov, Ludmil B, Gundem, Gunes, et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature*, 513(7518):422–425, 2014.
- [12] Birbrair, Alexander, Zhang, Tan, Wang, Zhong-Min, Messi, Maria Laura, Olson, John D, Mintz, Akiva, and Delbono, Osvaldo. Type-2 pericytes participate in normal and tumoral angiogenesis. *American Journal of Physiology-Cell Physiology*, 307(1):C25–C38, 2014.
- [13] Bishop, Christopher M. Pattern recognition. *Machine learning*, 128(9), 2006.
- [14] Burrell, Rebecca A and Swanton, Charles. Tumour heterogeneity and the evolution of poly-clonal drug resistance. *Molecular oncology*, 8(6):1095–1111, 2014.
- [15] Chen, Rui, Mias, George I, Li-Pook-Than, Jennifer, Jiang, Lihua, Lam, Hugo YK, Chen, Rong, Miriami, Elana, Karczewski, Konrad J, Hariharan, Manoj, Dewey, Frederick E, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 148(6):1293–1307, 2012.
- [16] Ciregan, Dan, Meier, Ueli, and Schmidhuber, Jürgen. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3642–3649. IEEE, 2012.
- [17] Cooper, Geoffrey M. *Elements of human cancer*. Jones & Bartlett Learning, 1992.
- [18] Davis, Alexander and Navin, Nicholas E. Computing tumor trees from single cells. *Genome biology*, 17(1):1–4, 2016.
- [19] de Visser, J Arjan GM and Rozen, Daniel E. Clonal interference and the periodic selection of new beneficial mutations in escherichia coli. *Genetics*, 172(4):2093–2100, 2006.
- [20] Demichelis, R, Retsky, MW, Hrushesky, WJM, Baum, M, and Gukas, ID. The effects of surgery on tumor growth: a century of investigations. *Annals of oncology*, 19(11):1821–1828, 2008.
- [21] Dentro, Stefan C, Leshchiner, Ignaty, Haase, Kerstin, Tarabichi, Maxime, Wintersinger, Jeff, Deshwar, Amit G, Yu, Kaixian, Rubanova, Yulia, Macintyre, Geoff, Vázquez-García, Ignacio, et al. Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. *BioRxiv*, page 312041, 2018.

- [22] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [23] Dhungel, Neeraj, Carneiro, Gustavo, and Bradley, Andrew P. Fully automated classification of mammograms using deep residual neural networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 310–314. IEEE, 2017.
- [24] Donmez, Nilgun, Malikic, Salem, Wyatt, Alexander W, Gleave, Martin E, Collins, Colin C, and Sahinalp, S Cenk. Clonality inference from single tumor samples using low coverage sequence data. In *International Conference on Research in Computational Molecular Biology*, pages 83–94. Springer, 2016.
- [25] Eaton, Jesse, Wang, Jingyi, and Schwartz, Russell. Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics*, 34(13):i357–i365, 2018.
- [26] El-Kebir, Mohammed. Sphyr: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics*, 34(17):i671–i679, 2018.
- [27] El-Kebir, Mohammed, Oesper, Layla, Acheson-Field, Hannah, and Raphael, Benjamin J. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70, 2015.
- [28] Fearon, Eric R and Vogelstein, Bert. A genetic model for colorectal tumorigenesis. *cell*, 61(5):759–767, 1990.
- [29] Fedele, Clare, Tothill, Richard W, and McArthur, Grant A. Navigating the challenge of tumor heterogeneity in cancer therapy. *Cancer discovery*, 4(2):146–148, 2014.
- [30] Fisher, Rosie, Pusztai, Lazos, and Swanton, C. Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, 108(3):479–485, 2013.
- [31] Friedl, Peter and Wolf, Katarina. Plasticity of cell migration: a multiscale tuning model. *Journal of Cell Biology*, 188(1):11–19, 2010.
- [32] Fukushima, Kunihiko. Neocognitron. *Scholarpedia*, 2(1):1717, 2007.
- [33] Gelman, Andrew, Shirley, Kenneth, et al. Inference from simulations and monitoring convergence. *Handbook of markov chain monte carlo*, 6:163–174, 2011.
- [34] Gerlinger, Marco, Rowan, Andrew J, Horswell, Stuart, Larkin, James, Endesfelder, David, Gronroos, Eva, Martinez, Pierre, Matthews, Nicholas, Stewart, Aengus, Tarpey, Patrick, et al.

Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl j Med*, 366:883–892, 2012.

- [35] Greaves, Mel and Maley, Carlo C. Clonal evolution in cancer. *Nature*, 481(7381):306–313, 2012.
- [36] Halford, S, Rowan, A, Sawyer, E, Talbot, I, and Tomlinson, Ian. O6-methylguanine methyltransferase in colorectal cancers: detection of mutations, loss of expression, and weak association with g: C> a: T transitions. *Gut*, 54(6):797–802, 2005.
- [37] Hanahan, Douglas and Weinberg, Robert A. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.
- [38] Hanahan, Douglas and Weinberg, Robert A. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.
- [39] Handa, Osamu, Naito, Yuji, and Yoshikawa, Toshikazu. Redox biology and gastric carcinogenesis: the role of helicobacter pylori. *Redox Report*, 16(1):1–7, 2011.
- [40] Hastings, W Keith. Monte carlo sampling methods using markov chains and their applications. 1970.
- [41] Hou, Yong, Song, Luting, Zhu, Ping, Zhang, Bo, Tao, Ye, Xu, Xun, Li, Fuqiang, Wu, Kui, Liang, Jie, Shao, Di, et al. Single-cell exome sequencing and monoclonal evolution of a jak2-negative myeloproliferative neoplasm. *Cell*, 148(5):873–885, 2012.
- [42] Hugo, Honor, Ackland, M Leigh, Blick, Tony, Lawrence, Mitchell G, Clements, Judith A, Williams, Elizabeth D, and Thompson, Erik W. Epithelial—mesenchymal and mesenchymal—epithelial transitions in carcinoma progression. *Journal of cellular physiology*, 213(2):374–383, 2007.
- [43] Husić, Edin, Li, Xinyue, Hujdurović, Ademir, Mehine, Miika, Rizzi, Romeo, Mäkinen, Veli, Milanič, Martin, and Tomescu, Alexandru I. Mipup: minimum perfect unmixed phylogenies for multi-sampled tumors via branchings and ilp. *Bioinformatics*, 35(5):769–777, 2019.
- [44] Jahn, Katharina, Kuipers, Jack, and Beerenwinkel, Niko. Tree inference for single-cell data. *Genome biology*, 17(1):1–17, 2016.
- [45] Kim, Kyung In and Simon, Richard. Using single cell sequencing data to model the evolutionary history of a tumor. *BMC bioinformatics*, 15(1):1–13, 2014.
- [46] LeCun, Yann, Bengio, Yoshua, and Hinton, Geoffrey. Deep learning. *nature*, 521(7553):436–444, 2015.

- [47] Lee, Kyung-Hwa, Lee, Ji-Shin, Nam, Jong-Hee, Choi, Chan, Lee, Min-Cheol, Park, Chang-Soo, Juhng, Sang-Woo, and Lee, Jae-Hyuk. Promoter methylation status of hmlh1, hmsh2, and mgmt genes in colorectal cancer associated with adenoma–carcinoma sequence. *Langenbeck's archives of surgery*, 396(7):1017–1026, 2011.
- [48] Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, and Stoyanov, Veselin. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [49] Malikic, Salem, Jahn, Katharina, Kuipers, Jack, Sahinalp, S Cenk, and Beerenwinkel, Niko. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nature communications*, 10(1):1–12, 2019.
- [50] Malikic, Salem, McPherson, Andrew W, Donmez, Nilgun, and Sahinalp, Cenk S. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(9):1349–1356, 2015.
- [51] Malikic, Salem, Mehrabadi, Farid Rashidi, Azer, Erfan Sadeqi, Ebrahimabadi, Mohammad Haghiri, and Sahinalp, S Cenk. Studying the history of tumor evolution from single-cell sequencing data by exploring the space of binary matrices. *bioRxiv*, 2020.
- [52] Malikic, Salem, Mehrabadi, Farid Rashidi, Ciccolella, Simone, Rahman, Md Khaledur, Ricketts, Camir, Haghshenas, Ehsan, Seidman, Daniel, Hach, Faraz, Hajirasouliha, Iman, and Sahinalp, S Cenk. Phiscs: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. *Genome research*, 29(11):1860–1877, 2019.
- [53] McGranahan, Nicholas and Swanton, Charles. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*, 168(4):613–628, 2017.
- [54] McPherson, Andrew, Roth, Andrew, Laks, Emma, Masud, Tehmina, Bashashati, Ali, Zhang, Allen W, Ha, Gavin, Biele, Justina, Yap, Damian, Wan, Adrian, et al. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature genetics*, 48(7):758, 2016.
- [55] Nik-Zainal, Serena, Van Loo, Peter, Wedge, David C, Alexandrov, Ludmil B, Greenman, Christopher D, Lau, King Wai, Raine, Keiran, Jones, David, Marshall, John, Ramakrishna, Manasa, et al. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, 2012.
- [56] Nowell, Peter C. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

- [57] Ricketts, Camir, Seidman, Daniel, Popic, Victoria, Hormozdiari, Fereydoun, Batzoglou, Serafim, and Hajirasouliha, Iman. Meltos: multi-sample tumor phylogeny reconstruction for structural variants. *Bioinformatics*, 36(4):1082–1090, 2020.
- [58] Ross, Edith M and Markowetz, Florian. Onconem: inferring tumor evolution from single-cell sequencing data. *Genome biology*, 17(1):1–14, 2016.
- [59] Sabeh, Farideh, Shimizu-Hirota, Ryoko, and Weiss, Stephen J. Protease-dependent versus-independent cancer cell invasion programs: three-dimensional amoeboid movement revisited. *Journal of Cell Biology*, 185(1):11–19, 2009.
- [60] Sadeqi Azer, Erfan, Rashidi Mehrabadi, Farid, Malikić, Salem, Li, Xuan Cindy, Bartok, Osnat, Litchfield, Kevin, Levy, Ronen, Samuels, Yardena, Schäffer, Alejandro A, Gertz, E Michael, et al. Phiscs-bnb: a fast branch and bound algorithm for the perfect tumor phylogeny reconstruction problem. *Bioinformatics*, 36(Supplement_1):i169–i176, 2020.
- [61] Sakr, WA, Haas, GP, Cassin, BF, Pontes, JE, and Crissman, JD. The frequency of carcinoma and intraepithelial neoplasia of the prostate in young male patients. *The Journal of urology*, 150(2):379–385, 1993.
- [62] Salehi, Sohrab, Steif, Adi, Roth, Andrew, Aparicio, Samuel, Bouchard-Côté, Alexandre, and Shah, Sohrab P. ddclone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome biology*, 18(1):1–18, 2017.
- [63] Satas, Gryte and Raphael, Benjamin J. Tumor phylogeny inference using tree-constrained importance sampling. *Bioinformatics*, 33(14):i152–i160, 2017.
- [64] Satas, Gryte, Zaccaria, Simone, Mon, Geoffrey, and Raphael, Benjamin J. Scarlet: Single-cell tumor phylogeny inference with copy-number constrained mutation losses. *Cell Systems*, 10(4):323–332, 2020.
- [65] Selsam, Daniel, Lamm, Matthew, Bünz, Benedikt, Liang, Percy, de Moura, Leonardo, and Dill, David L. Learning a sat solver from single-bit supervision. *arXiv preprint arXiv:1802.03685*, 2018.
- [66] Senior, Andrew W, Evans, Richard, Jumper, John, Kirkpatrick, James, Sifre, Laurent, Green, Tim, Qin, Chongli, Žídek, Augustin, Nelson, Alexander WR, Bridgland, Alex, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

- [67] Silver, David, Schrittwieser, Julian, Simonyan, Karen, Antonoglou, Ioannis, Huang, Aja, Guez, Arthur, Hubert, Thomas, Baker, Lucas, Lai, Matthew, Bolton, Adrian, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.
- [68] Singer, Jochen, Kuipers, Jack, Jahn, Katharina, and Beerenwinkel, Niko. Single-cell mutation identification via phylogenetic inference. *Nature communications*, 9(1):1–8, 2018.
- [69] Sokal, Alan. Monte carlo methods in statistical mechanics: foundations and new algorithms. In *Functional integration*, pages 131–192. Springer, 1997.
- [70] Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
- [71] Stewart, BWKP and Wild, CP. World cancer report 2014. health, 2017.
- [72] Stratton, Michael R, Campbell, Peter J, and Futreal, P Andrew. The cancer genome. *Nature*, 458(7239):719–724, 2009.
- [73] Strino, Francesco, Parisi, Fabio, Micsinai, Mariann, and Kluger, Yuval. Trap: a tree approach for fingerprinting subclonal tumor composition. *Nucleic acids research*, 41(17):e165–e165, 2013.
- [74] Sun, Xiao-xiao and Yu, Qiang. Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. *Acta Pharmacologica Sinica*, 36(10):1219–1227, 2015.
- [75] Sutherland, NS. Outlines of a theory of visual pattern recognition in animals and man. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 171(1024):297–317, 1968.
- [76] Talbot, Simon J and Crawford, Dorothy H. Viruses and tumours—an update. *European Journal of Cancer*, 40(13):1998–2005, 2004.
- [77] Truninger, Kaspar, Menigatti, Mirco, Luz, Judith, Russell, Anna, Haider, Ritva, Gebbers, Jan-Olaf, Bannwart, Fridolin, Yurtsever, Huseyin, Neuweiler, Joerg, Riehle, Hans-Martin, et al. Immunohistochemical analysis reveals high frequency of pms2 defects in colorectal cancer. *Gastroenterology*, 128(5):1160–1171, 2005.
- [78] Vander Heiden, Matthew G, Cantley, Lewis C, and Thompson, Craig B. Understanding the warburg effect: the metabolic requirements of cell proliferation. *science*, 324(5930):1029–1033, 2009.

- [79] Waclaw, Bartlomiej, Bozic, Ivana, Pittman, Meredith E, Hruban, Ralph H, Vogelstein, Bert, and Nowak, Martin A. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature*, 525(7568):261–264, 2015.
- [80] Williams, Ronald J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.
- [81] Wu, Yufeng. Accurate and efficient cell lineage tree inference from noisy single cell data: the maximum likelihood perfect phylogeny approach. *Bioinformatics*, 36(3):742–750, 2020.
- [82] Yuan, Ke, Sakoparnig, Thomas, Markowetz, Florian, and Beerenwinkel, Niko. Bitphylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome biology*, 16(1):1–16, 2015.
- [83] Zaccaria, Simone, El-Kebir, Mohammed, Klau, Gunnar W, and Raphael, Benjamin J. The copy-number tree mixture deconvolution problem and applications to multi-sample bulk sequencing tumor data. In *International Conference on Research in Computational Molecular Biology*, pages 318–335. Springer, 2017.
- [84] Zafar, Hamim, Navin, Nicholas, Chen, Ken, and Nakhleh, Luay. Sicleonefit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome research*, 29(11):1847–1859, 2019.
- [85] Zafar, Hamim, Tzen, Anthony, Navin, Nicholas, Chen, Ken, and Nakhleh, Luay. Sifit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome biology*, 18(1):1–20, 2017.
- [86] Zhu, Aizhi, Lee, Daniel, and Shim, Hyunsuk. Metabolic positron emission tomography imaging in cancer detection and therapy response. In *Seminars in oncology*, volume 38, pages 55–69. Elsevier, 2011.