فصل ۱

روش‌های پیشین

۱.۱ 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.

۲.۱ مدل کیم و سایمون[۱۳]

این مدل در سال ۲۰۱۴ با تمرکز بر ساخت درخت فیلوژنی[1] از طریق رابطه ترکیبی میان جهش‌های ایجاد شده در داده‌های توالی‌یابی تک‌سلولی دی‌ان‌ای[2] 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

[1] Phylogeny tree
[2] DNA

جدول ۱.۱: 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

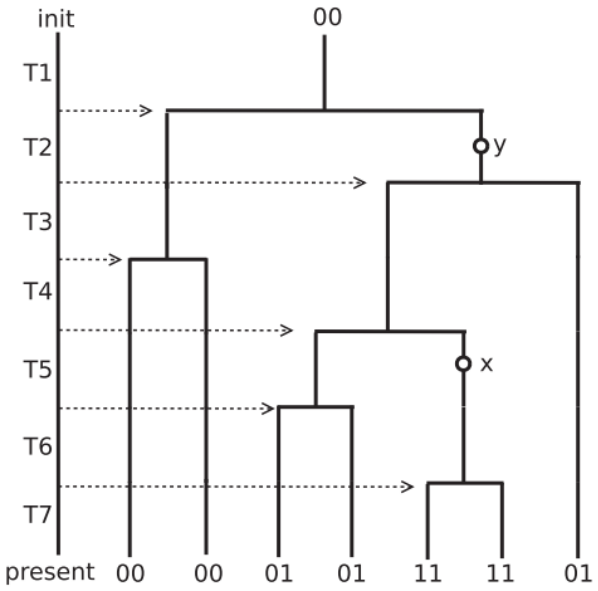
توالی‌یابی تک‌سلولی تا نزدیک‌ترین جد مشترک وجود دارد. استنباط درخت فیلوژنی از طریق لگوریتم کیم و سایمون، بر مبنای منطق بیزی است، یعنی از این منطق به منظور تعیین رابطه ترتیبی بین هر دو جهش گوناگون استفاده شده است. در ادامه مقدار بیشینه درست‌نمایی[۳] درخت استنباط شده بر مبنای احتمال ترتیبی دوبه‌دوی بین 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 توالی‌یابی و ترک آلل[۴] 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 ۱.۱ نشان داده شده‌است. در این جدول صفر بیانگر عدم وجود جهش و یک بیانگر وجود جهش است. تعداد رخداد جهش‌ها با فرض عدم وجود خطا در توالی‌یابی 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 (۰،۱) 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 استنباط کرد که y نسبت به x قدیمی‌تر است و به عنوان یکی از اجداد x در درخت فیلوژنی تومور قرار می‌گیرد. در نتیجه با استفاده از جدول داده‌های نمونه‌برداری شده، استنباط یک رابطه زمانی میان جهش‌های صورت گرفته امکان پذیر است.

شکل ۱.۱ یک درخت فیلوژنیک تومور را نشان می‌دهد که از داده‌های جدول بالا استنباط 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

[3] Maximum-likelihood
[4] Allele dropout

هیچ جهشی رخ نداده است. محور عمودی بیانگر سیر زمانی تکامل تومور است که به تعداد نمونه‌ها تقسیم شده است. برای استنباط درخت فیلوژنی تومور، الگوریتم کیم و سایمون از سه بخش اصلی تشکیل شده است. طبق



شکل ۱.۱: نمایی از یک درخت فیلوژنیک تومور

قضیه بیز برای محاسبه هر یک از این سه احتمال به مقادیر درست‌نمایی[۵] نیاز داریم. مقدار احتمال رخداد طبق رابطه زیر محاسبه می‌گردد:

$$P(x \sim y|D) \propto L(x \sim y)P(x \sim y), \quad L(x \sim y) = P(D|x \sim y) \tag{۱.۱}$$

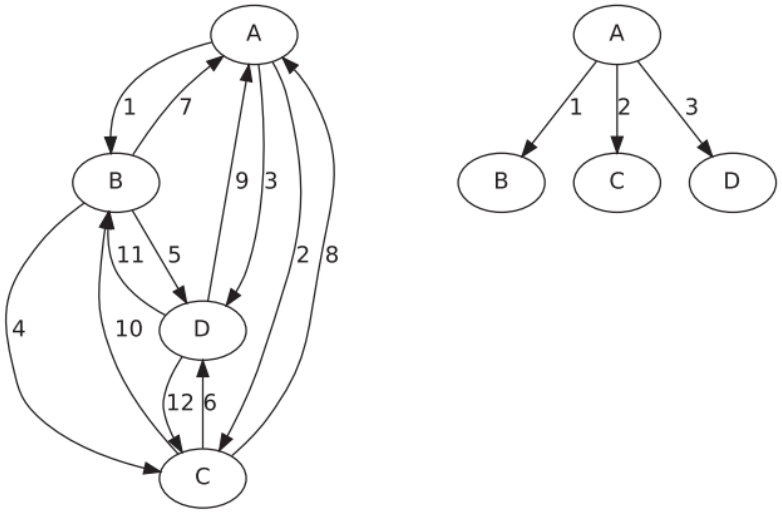
طبق این رابطه و با توجه به اینکه رابطه زمانی میان جهش‌های x و y دارای ۳ حالت،

$$x \to y, \qquad y \to x \qquad \text{و} \qquad x \not\leftrightarrow y$$

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 فیلوژنی قرار خواهد گرفت. طبق آنچه گفته شد یک گراف جهت دار فیلوژنی بلقوه مشابه آنچه در شکل ۲.۱ نشان داده شده است استنباط خواهد شد. در نهایت از این گراف جهت دار، یک درخت

[5] Likelihood

به طوری که روابط میان جهش‌ها از آن استنباط شود ساخته خواهد شد. در ابتدا یال‌های گراف از طریق رابطه‌ای



شکل ۲.۱: یک گراف جهت دار فیلوژنی

که در ادامه آمده است وزن‌دهی می‌شوند،

$$w_{x\sim y} = -\log P(x \sim y|D) \tag{۲.۱}$$

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.

$$\hat{T} = \arg\min\Big(\sum_{x\sim y\in T} w_{x\sim y}\Big) = \arg\max\Big(\prod_{x\sim y\in T} P(x \sim y|D)\Big) \tag{۳.۱}$$

در شکل ۲.۱ محتمل‌ترین درخت فیلوژنی 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 کمینه سازی برابر هفت است که این مقدار کمترین مقدار ممکن است.

۱.۲.۱ پایگاه داده

در این مقاله از پایگاه داده تولی‌یابی تک سلولی هو و همکاران [۱۰] استفاده شده است. این مجموع داده از توالی‌یابی تک‌سلولی دی‌ان‌ای نمونه‌های توموری یک نوع خاص از سرطان خون[۶] 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 ژنوتایپی 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 فیلوژنی نسبت داده شده به این پایگاه داده زیاد باشد. در واقع با در نظر گرفتن حالت‌های مختلف روابط دوبه‌دوی جهش‌های گوناگون، می‌توان درخت‌های جهشی متنوعی از داده‌ها استنباط کرد.

۲.۲.۱ معیار ارزیابی

ارزیابی درختهای جهشی گوناگون از طریق روش LOOCV[7] صورت میگیرد. این روش همانند روش اارزیابی‌های متقابل[۸] با K قسمت می‌باشد با این تفاوت که در آن k برابر تعداد جهش‌ها (تعداد ستون‌های ماتریس ژنوتایپ) 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.

[6] Thrombocythemia
[7] Leave one out cross validation
[8] Cross validation

۳.۱ الگوریتم Bitphylogeny [۳۴]

این الگوریتم در سال ۲۰۱۵ ارئه 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 انبوه[۹] برای استنباط درخت فیلوژنی 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 ژنوتایپی) این الگوریتم تعدادی سطر و ستون است که در آن سطرها بیانگر سلولها و ستونها نمایانگر انواع جهشهای مختلف است. این ماتریس، یک ماتریس دودویی‌ها[۱۰] است که در آن بودن درایه i و j بیانگر آن است که در سطر iام جهشی از نوع j ام وجود ندارد. متعاقباً، اگر مقدار درایه i و j برابر یک باشد در سطر iام جهشی از نوع jام وجود دارد.

دراین مقاله برای جستجو درختی که بیشترین تطابق با داده‌های ورودی را داشته باشد از الگوریتم زنجیره مارکوف مونت کارلو[۱۱] استفاده می‌شود. این الگوریتم سلول‌ها با ژنوتایپ 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 فیلوژنی به سطوح پایین‌تر را محاسبه کند. در این الگوریتم از داده‌های تغییرات تک نوکلئوتید[۱۲] 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

[9] Bulk sequencing
[10] Binary
[11] Markov Chain Monte Carlo (MCMC)
[12] Nucleotid

هر جهش در درخت فیلوژنی تومور اشاره کرد اما این مقدار احتمال بدلیل تعداد بالای داده‌های از دست رفته و نرخ بالای خطای مثبت کاذب[۱۳] و منفی کاذب[۱۴]، 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 فیلوژنی 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.

۱.۳.۱ پایگاه داده

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

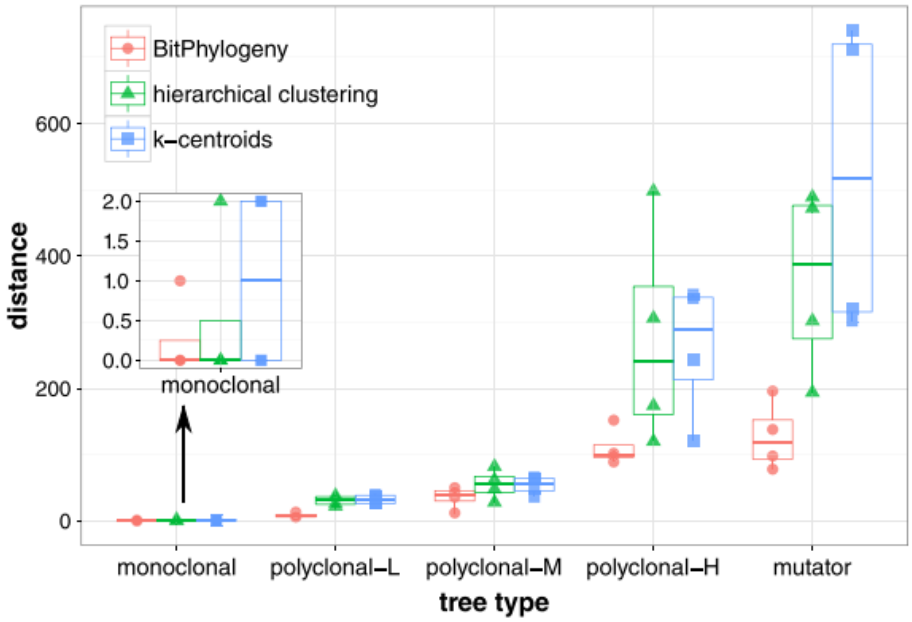
- دادگان مربوط به الگوهای متیلاسیون سرطان روده بزرگ
- دادگان شبیه‌سازی شده مربوط به سرطان خون

۲.۳.۱ معیار ارزیابی

به منظور ارزیابی عمکرد الگوریتم بیت‌فیلوژنی یک مقایسه بین خروجی این الگوریتم و خروجی‌های الگوریتم‌های خوشه‌بندیk هسته‌ای[۱۵] و دسته‌بندی سلسله مراتبی[۱۶] 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

[13] False positive
[14] False negative
[15] K-Centroids
[16] Hierarchical clustering

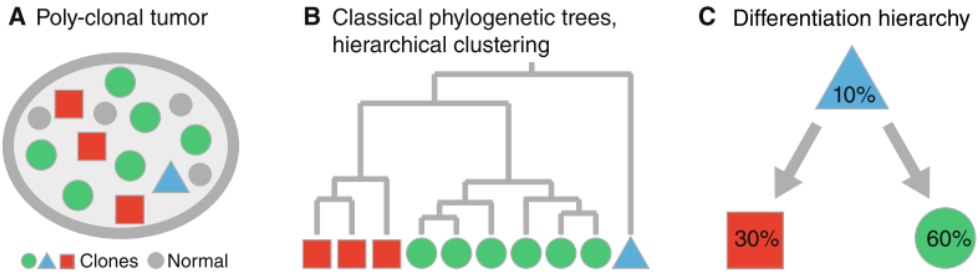
مقاله گواه از پایداری[۱۷] و دقت[۱۸] بسیار بهتر الگریتم بیت‌فیلوژنی نسبت به دو الگوریتم دیگر است. در شکل ۳.۱ میزان خطای عملکرد الگوریتم بیت‌فیلوژنی نسبت به دو الگوریتم خوشه‌بندیk هسته‌ای و دسته‌بندی سلسله مراتبی در سطوح مختلف درخت در حالت‌های تک‌کلونی و چندکلونی قابل مشاهده است.



شکل ۳.۱: میزان خطای عملکرد الگوریتم بیت‌فیلوژنی نسبت به دو الگوریتم خوشه‌بندیk هسته‌ای و دسته‌بندی سلسله مراتبی در سطوح مختلف درخت در حالت‌های تک‌کلونی و چندکلونی [۳۴]

در شکل ۴.۱ به طور کلی مراحل عمکلرد الگوریتم بیت‌فیلوژنی 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 بیت‌فیلوژنی مشاهده می‌شود که در آن کلون‌ها و فراوانی هر یک مشهود است.

[17] Consistency
[18] Accuracy



شکل ۴.۱: مراحل عملکرد الگوریتم بیت‌فیلوژنی

۴.۱ الگوریتم SCITE [۱۲]

این الگوریتم با استفاده از داده‌های توالی‌یابی تک سلولی[۱۹] سعی در استنباط درخت فیلوژنی 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 منحصربه‌فرد را از طریق تکثیر سلولی به سلول‌های بعدی منتقل کنند. [۴]

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. [۲]

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 [۱۹]. 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. [۲۴]

الگوریتم Scite از طریق داده‌های توالی‌یابی تک سلولی قادر است سیر تکاملی تومور را از طریق درخت جهشی تومور که درآن ترتیب وقوع جهش‌ها مشخص است یا از طریق استنباط درخت فیلوژنیک تومور که در آن هر برگ نشان دهنده یک سلول است، نشان دهد. خروجی مدل Scite نتیجه ارزیابی بهتری در مقایسه با

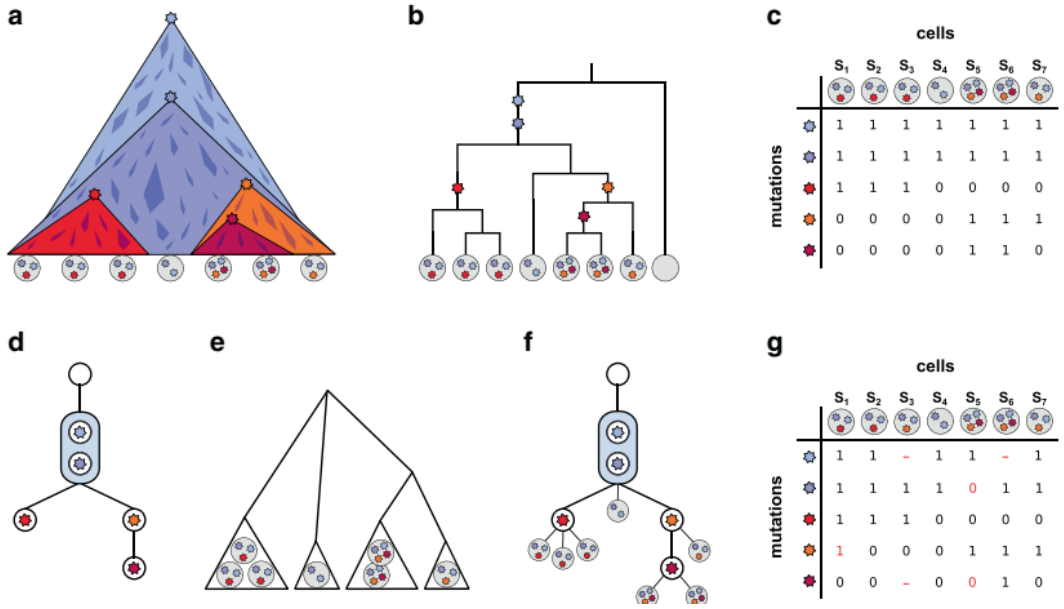
[19] Single cell sequencing

الگوریتم بیت‌فیلوژنی بر روی داده‌های واقعی داراست. الگوریتم Scite از طریق معیار بیشینه درست‌نمایی و احتمال رخداد هر جهش و با استفاده از ماتریس ژنوتایپ 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 ژنوتایپ ورودی زیاد باشد، ساخت درخت فیلوژنیک راحت‌تر خواهد بود، اما در حالتی که تعداد سلولها زیاد باشد (تعداد سطرهای ماتریس ژنوتایپ بالا باشد) ساخت درخت جهشی تومور (ترتیب وقوع جهش‌ها) راحت‌تر است. به طور خلاصه اینکه کدام نوع درخت (جهشی یا فیلوژنیک) در نهایت بیان‌کننده سیر تکاملی تومور باشد به نرخ جهش‌های توموری و روش توالی‌یابی داده‌ها بستگی دارد.
در الگوریتم Scite از دو فرض اصلی استفاده می‌شود:

- فرض مکان‌های بی‌نهایت[۲۰] 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 بیت‌فیلوژنی از یک ماتریس ژنوتایپ (ماتریس دودویی‌ها که سطرها نمایانگر نمونه‌ها و ستون‌ها بیانگر جهش‌هاست) به عنوان ورودی الگریتم استفاده می‌شود. موقعیت هر جهش به صورت درایه i و j از ماتریس $n * m$ مشخص می‌شود. به این صورت که مقدار صفر در سطر iام و درایه jام بیانگر آن است که جهش از نوع j در سطر i وجود ندارد. یک ماتریس ژنوتایپ را ماتریس فیلوژنی کامل[۲۱] گوییم هر گاه به ازای آن یک درخت فیلوژنیک متناظر باشد. در الگوریتم scite همانند الگوریتم بیت‌فیلوژنی از الگوریتم زنجیره مارکوف مونت کارلو برای استنباط درخت تتکاملی تومور از داده‌های توالی‌یابی تک سلولی استفاده می‌شود، با این تفاوت که فضای جستجو برای انتخاب پارامترها بسیار محدودتر از حالت بیت‌فیلوژنی است و نرخ خطاهای داده (مثبت کاذب و منفی کاذب) برای همه جهش‌ها یکسان در نظر گرفته شده است. محدود کردن فضا جستجو برای انتخاب پارامترها از طریق نمونه‌برداری در این فضا، 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 خطا توالی‌یابی است. شکل ۵.۱ یک استنتاج تکاملی از داده‌های توالی‌یابی تک سلولی را نشان می‌دهد. در این شکل، a یک ماتریس فیلوژنی 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 فیلوژنی کامل را با E نشان داده شده است.

[20] Infinite sites
[21] Perfect phylogenic matrix



شکل ۵.۱: یک استنتاج تکاملی از داده‌های توالی‌یابی تک سلولی [۱۲]

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 ۴.۱ تعریف می‌گردند.

$$\begin{aligned} P(D_{ij} = ۱|E_{ij} = ۰) = \alpha, \qquad & P(D_{ij} = ۰|E_{ij} = ۰) = ۱ - \alpha \\ P(D_{ij} = ۰|E_{ij} = ۱) = \beta, \qquad & P(D_{ij} = ۱|E_{ij} = ۱) = ۱ - \beta \end{aligned} \tag{۴.۱}$$

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

$$y = xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx \tag{۵.۱}$$

در معادله بالا E ماتریس جهش‌دار است که با درخت جهشی T و بردار ضمیمه ∂ تعریف می‌گردد. توزیع

احتمال پسین به صورت زیر محاسبه می‌گردد:

$$y = xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx \quad (۶.۱)$$

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

$$y = xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx \quad (۷.۱)$$

اندازه فضای جستجو برای دو پارامتر درخت جهشی T و بردار ضمیمه ∂ برابر با $(n+1)^{n-1} \times (n+1)^m$ می‌باشد. این فضا جستجو با فرض یکنواخت بودن توزیع بردار ضمیمه ∂ و طبق معادله بالا و حذف بردار ضمیمه به $(n+1)^{n-1}$ 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:

$$y = xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx \quad (۸.۱)$$

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

۱.۴.۱ پایگاه داده

به منظور ارزیابی عملکرد الگوریتم Scite برای استنباط 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 $۶/۴ \times ۱۰^{-۴}$ و نرخ خطا منفی کاذب ۰/۴۳۰۹ .

- داده‌های توالی‌یابی تک سلولی یک نوع خاص از سرطان کبد با ۱۷ سلول سرطانی و ۵۰ نوع جهش با مقادیر نرخ خطای مثبت کاذب $۲/۶۷ \times ۱۰^{-۵}$ و نرخ خطا منفی کاذب ۰/۱۶۴۳ و نرخ داده‌های از دست رفته ۲۲ درصد.

- داده‌های توالی‌یابی تک سلولی‌نمونه‌گیری شده از سرطان سینه با ۴۷ سلول سرطانی و ۴۰ نوع جهش و با نرخ خطای ترک آلل ۷۳ درصد و نرخ خطای مثبت کاذب $۱/۲۴ \times ۱۰^{-۶}$.

شایان ذکر است که مدت زمان استنباط یک درخت فیلوژنی 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 رخ دهد 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.

۵.۱ الگوریتم Onconem [۲۲]

این الگوریتم در سال ۲۰۱۶ با هدف یافتن تاریخچه تکاملی ناحیه‌های درون توموری با استفاده از داده‌های توالی‌یابی تک سلولی ارائه گردید. این الگوریتم قادر است تا ناحیه‌های درون توموری مشابه را درون یک دسته قرار دهد و برای آنها یک ژنوتایپ یکتا در نظر بگیرد. این الگوریتم بر مبنای تغییرات تک نوکلئوتیدی، درخت تکاملی تومور را استنباط می‌کند و قادر به یافتن خطاهای ژنوتایپی 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 انتها رابطه میان جهش‌ها و ژنوتایپ‌های مشاهده شده و مشاهده نشده (پیش‌بینی شده) مشخص می‌گردد. این الگوریتم هم می‌تواند درخت کلونال توموری و هم درخت فیلوژنیک توموری (قرارگرفتن سلول‌ها به عنوان برگهای درخت) را به عنوان خروجی بدست دهد. ورودی این الگوریتم ماتریس دودویی ژنوتایپ به همراه نرخ خطا مثبت کاذب و نرخ خطا منفی کاذب و نرخ خطا داده‌های از دست رفته است. در ادامه، الگوریتم سعی می‌کند تا سلول‌ها با

ژنوتایپ‌های 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 جمیع 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 داده‌های مدل به صورت زیر محاسبه می‌گردد

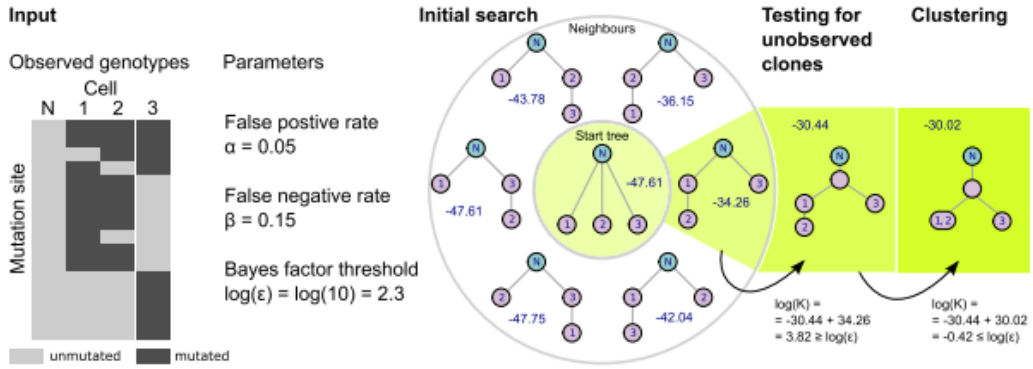
$$y = xxxxxxxxxxxxxxxxxxxxxxxx \tag{۹.۱}$$

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 $p(\tau)$ دارای توزیع یکنواخت است. رابطه بالا می‌تواند به شکل زیر بازنویسی شود

$$y = xxxxxxxxxxxxxxxxxxxxxxxx \tag{۱۰.۱}$$

بر طبق این رابطه، برای درختی با n راس، فضا جستجو شامل $n^{n-۲}$ انتخاب است که هزینه محاسباتی بسیار بالایی برای درختانی با راس‌های بیشتر از ۹ دارد. طبق شکل ۶.۱، برای محدود کردن فضای جستجو از یک الگوریتم اکتشافی استفاده می‌شود تا اطمینان حاصل شود که خروجی الگوریتم یک نقطه بهینه محلی 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.

روند کلی الگوریتم Onconem در شکل بالا توضیح داده است. طبق این شکل، ماتریس دودویی ژنوتایپی به همراه نرخ خطاهای α و β به عنوان ورودی الگوریتم استفاده می‌شوند. طبق شکل بالا میزان درست‌نمایی اولیه برابر $-۴۷/۶۱$ محاسبه شده است اما از میان همه درخت‌های همسایه درخت اولیه، آن درختی که بیشترین



شکل ۱.۶: نمای شماتیکی از الگوریتم Onconem [۲۲]

درست‌نمایی را دارد به عنوان درخت اولیه انتخاب می‌شود (با درست‌نمایی ۳۶٫۲۶−). 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 ۳٫۸۲ افزایش می‌یابد و این کلون مشاهده نشده بدلیل بزرگتر بودن مقدار درست‌نمایی از آستانه تعیین شده، به مدل افزوده می‌شود. در نهایت گره‌های یک شاخه تا جایی که سبب کاهش میزان درست‌نمایی نشوند، در یک کلون تجمیع می‌شوند.

۱.۵.۱ پایگاه داده

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

- داده‌های توالی‌یابی 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[۲۲] اند . خروجی الگوریتم Onconem برای این پایگاه داده یک درخت فیلوژنی 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 بیت‌فیلوژنی 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 بیت‌فیلوژنی به آن رسیده بود (۱۱۵۸۴−).

[22] Missing data

۶.۱ الگوریتم Sasc [۲۲]

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، خصویات 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 تومور فرض‌های گوناگونی جهت ساده‌سازی مسئله صورت می‌گیرد، مثل فرض مکان‌های بی‌نهایت که طبق آن هر جهش یکتایی تنها یکبار رخ می‌دهد. مطالعات زیادی صورت گرفته است که نشان می‌دهد در نظر گرفتن فرض مکان‌های بی‌نهایت به تنهایی 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 راه‌حل 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 مثبت کاذب، نرخ خطای منفی کاذب و نرخ خطای اطلاعات از دست رفته است و بیشینه درست‌نمایی از رابطه زیر بدست می‌آید:

$$y = xxxxxxxxxxxxxxxxxxxxxxxxxx \tag{۱۱.۱}$$

۱.۶.۱ پایگاه داده:

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

- داده‌های توالی‌یابی تک سلولی سرطان مثانه
- داده‌های شبیه‌سازی شده سرطان خون

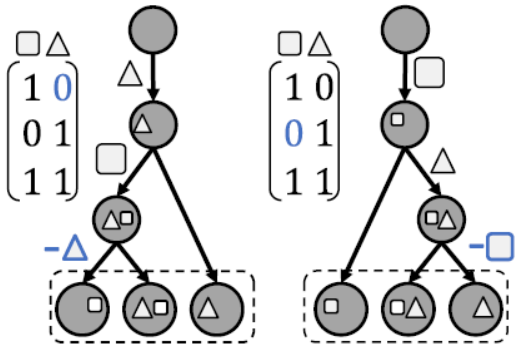
خروجی الگوریتم در مقایسه با الگوریتم Scite از مقدار بیشینه درست‌نمایی بیشتری برای مدل کردن داده‌ها برخوردار است.

۷.۱ الگوریتم Scarlet [۲۶]

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 خطا[۲۳] در نظر می‌گیرد. این مدل اجازه حذف دگرگونی تک‌هسته‌ای[۲۴] را تنها هنگامی که با شواهد داده توالی‌یابی تک سلولی دی‌ان‌ای از یک حذف در همان مکان هندسی[۲۵] همراه شده باشد، می‌دهد. این مدل پایه الگوریتم اسکارلت خواهد بود که فیلوژنی تومور را از داده توالی یابی تک سلولی 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 فیلوژنی داده توالی‌یابی تک سلولی 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، فیلوژنی تومور را از داده توالی‌یابی تک سلولی 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 تک سلولی دی‌ان‌ای از تومور بدلیل افزایش بازدهی الگوریتم و کاهش هزینه ایزوله کردن، نشانه‌گذاری و توالی‌یابی سلول‌های انفرادی از محبوبیت روزافزونی برخوردار است.

[23] loss-supported
[24] Single nucleotide variant (SNV)
[25] Loci

شکل ۷.۱ فیلوژنی با فرض دولو[۲۶] را نشان می‌دهد. این مدل با شناسایی حذف جهش‌ها به منظور رفع تناقض مدل مکان‌های بی‌نهایت می‌تواند درخت‌های ۷.۱ را بسازد. هر دو مدل دولو و مکان‌های بی‌نهایت می‌توانند چندین درخت ممکن را بسازند. حتی در حالت های ساده ای که خطا وجود ندارد، استنباط چندین فیلوژنی



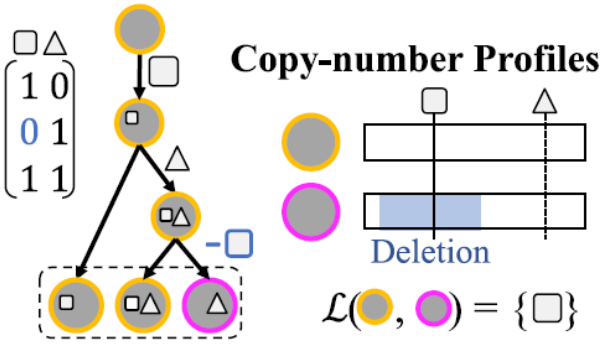
شکل ۷.۱: عنوانننننننننننننننننننننننننننننننن

سازگار با داده‌ها ممکن است وجود داشته باشند. در صورتی که خطا وجود داشته باشد و عدم قطعیت در ماتریس جهش وجود داشته باشد، تعداد این درخت‌های احتمالی بسیار بیشتر خواهد شد. خطای داده توالی‌یابی تک سلولی دی‌ان‌ای و حذف جهش‌ها منجر به پیچیدگی مسئله و ابهام در استنباط فیلوژنی خواهد شد. به عنوان مثال با مشاهده کردن ۰ در ماتریس جهش به جای ۱ نمی‌توان براحتی بین خطاهای داده‌ها و حذف جهش‌ها تفاوتی قائل شد. عمده 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[۲۷] 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 صدها 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[۲۸]،

[26] Dollo
[27] Multi-state
[28] Copy number variation (CNV)

درخت فیلوژنی را براساس داده توالی‌یابی تک سلولی دی‌ان‌ای می‌سازد. الگوریتم اسکارلت براساس مدل فیلوژنی با در نظر گرفتن حذف خطا 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 فیلوژنی با در نظر گرفتن خطا حذف را نشان می‌دهد که با استفاده از داده تغییر تعداد کپی سعی در محدود کردن حذف جهش ها دارد تا بتواند ابهام[۲۹] ایجاد شده را رفع کند.



شکل ۸.۱: عنوانننننننننننننننننننننننننننننن

zzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzz

مدل فیلوژنی با در نظر گرفتن حذف خطا، مدلی از تکامل دگرگونی تک‌هسته‌ای است که جهش حداکثر یکبار رخ خواهد داد (a۱-o) اما حذف جهش‌ها (a۱–o) توسط مجموعه از مقدار خطا حذف که توسط تغییر تعداد کپی‌ها تعریف می‌شوند محدود خواهد شد. برای هر جفت سلول ، از مجموعه جهش‌های تغییرات تعداد کپی، مجموعه خطا به صورت ، تعریف خواهد شد. مدل فیلوژنی با در نظر گرفتن حذف خطا، توسعه‌دهنده مدل‌های مکان‌های بی‌نهایت و دولو می‌باشد. ضمنا الگوریتم اسکارلت متکی بر مدل احتمالاتی تعداد خوانش‌ها برای هر دگرگونی تک‌هسته‌ای است تا خطاها و داده‌های از بین‌رفته، که در توالی‌یابی تک سلولی دی‌ان‌ای معمول هستند، را مورد توجه قرار می‌دهد.

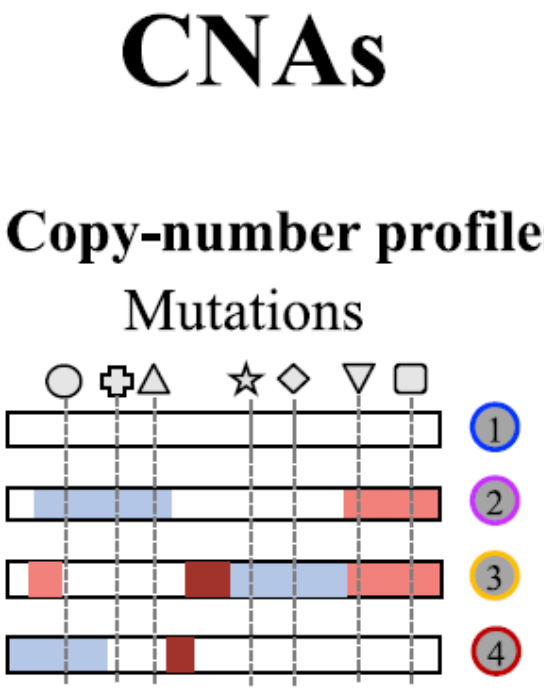
اسکارلت سه ویژگی مهم دارد:

- مدل فیلوژنی با در نظر گرفتن حذف خطا، حذف جهش ها را محدود به مکان‌هایی می‌کند که کاهش متناظر با آن در تعداد جهش‌های کپی وجود داشته باشد.

[29] conflict

- این الگوریتم با استفاده از مدل فیلوژنی با در نظر گرفتن حذف خطا، ابتدا درخت فیلوژنی اولیه استنتاج شده را پایش و سپس از طریق داده‌های تغییر تعداد کپی، فیلوژنی نهایی را استنباط می‌کند.
- استنتاج مبتنی بر بیشینه درست‌نمایی از دگرگونی‌های تک‌هسته‌ای با استفاده از مدل احتمالاتی تعداد خوانش‌های مشاهده شده در داده‌های توالی‌یابی تک سلولی دی‌ان‌ای توسط الگوریتم اسکارلت اجرا می‌شود.

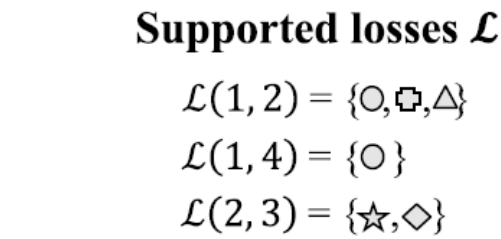
اگر تعدادی جهش‌های حذف و تغییر تعداد کپی موجود باشند و وضعیت جهش‌های تغییر تعداد کپی را با رنگ قرمز و حذف نواحی ژنوم در طول کل ژنوم را با رنگ آبی نشان دهیم:



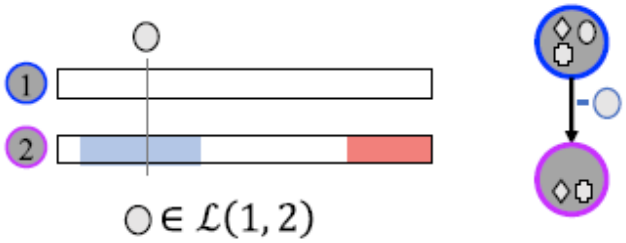
شکل ۹.۱: عنوانننننننننننننننننننننننننن

آنگاه مجموعه خطا مدل فیلوژنی با در نظر گرفتن حذف خطا، توسط مجموعه‌های ۱۱.۱ نمایش داده خواهد شد:

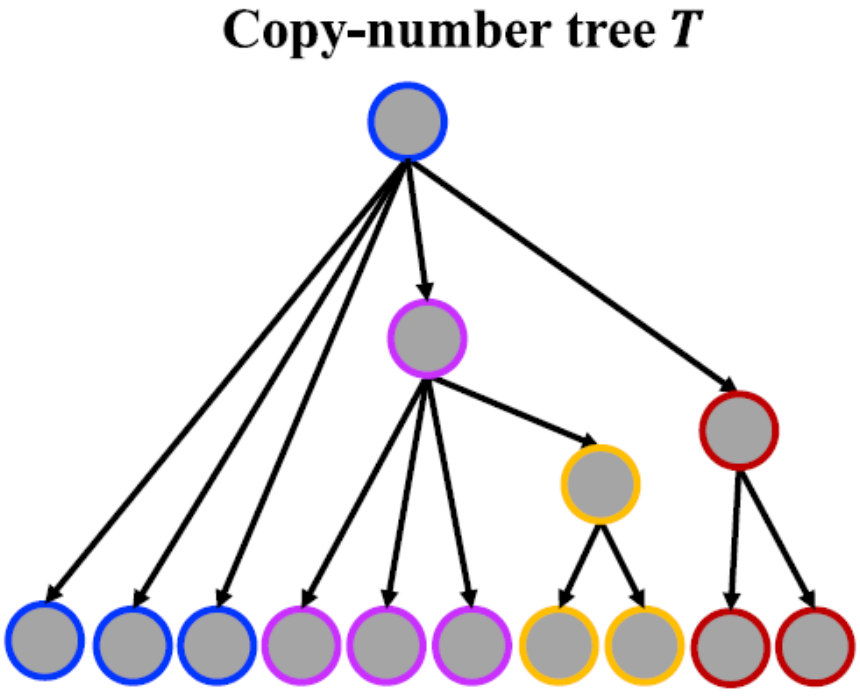
الگوریتم اسکارلت به صورت مستقیم وضعیت جهش‌های حذف و تغییر تعداد کپی سلول‌های پدری را نشان نخواهد داد. به منظور غلبه بر این موضوع یک درخت برای جهش‌های حذف و تغییر تعداد کپی زیر را در نظر خواهد گرفت که از روی وضعیت جهش‌های حذف و تغییر تعداد کپی سلول‌های مشاهده شده، ساخته شده است.



شکل ۱.۱۰: عنوانننننننننننننننننننننننن



شکل ۱.۱۱: عنوانننننننننننننننننننننننن

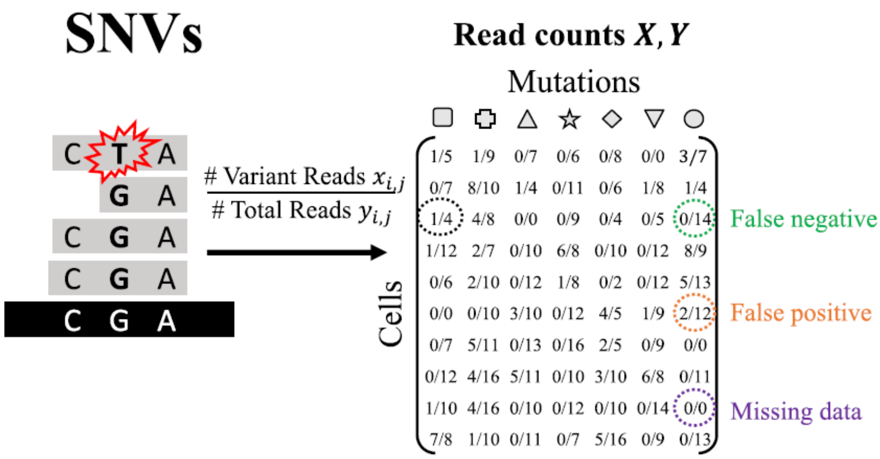


شکل ۱.۱۲: عنوانننننننننننننننننننننننن

دو ورودی برای الگوریتم اسکارلت در نظر گرفته می‌شود:

- 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 فیلوژنی 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 تنوع[۳۰] X و مجموع[۳۱] Y از تعداد خوانش‌ها[۳۲] برای هر سلول و هر جهش مطابق ماتریس ۱۳.۱ تهیه شده است:



شکل ۱۳.۱: عنوانننننننننننننننننننننننننننن

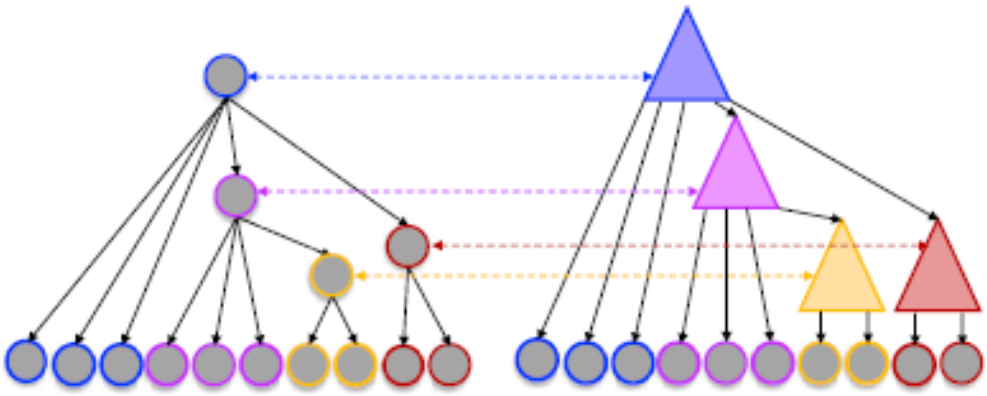
در ادامه الگوریتم اسکارلت، روابط بین اتصال سلولها (T') را از سلول های مشاهده شده(برگها) و ماتریس جهش بیشینه درست‌نمایی B^* را با محدود کردن حذف جهش ها به مجموعه

zzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzz

از خطاهای احتمالی حساب میکند. سپس با مقایسه T' از T و انتخاب بیشینه درست‌نمایی B^* را با استفاده از مدل احتمالاتی برای حضور ($b_{i,j} = ۱$) و یا عدم حضور ($b_{i,j} = ۰$) هردگرگونی تک‌هسته‌ای در هر سلول را انجام می‌دهد.

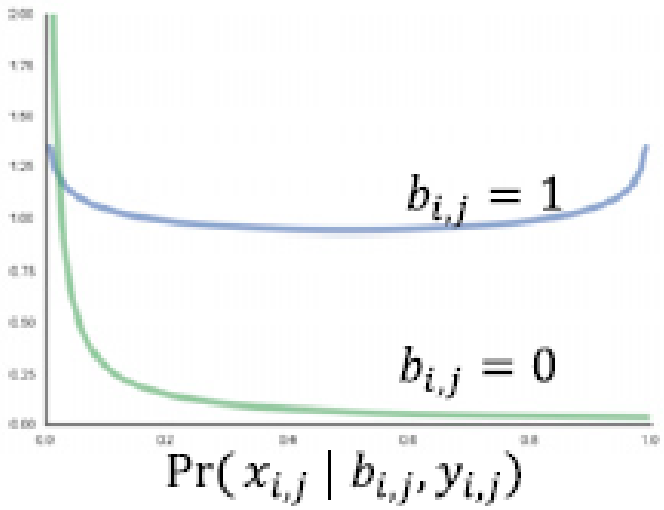
مقایسه T' از T:

[30] Variant
[31] Total
[32] Read counts



شکل ۱۴.۱: مقایسه T' از T

مدل احتمالاتی برای توالی‌یابی داده:

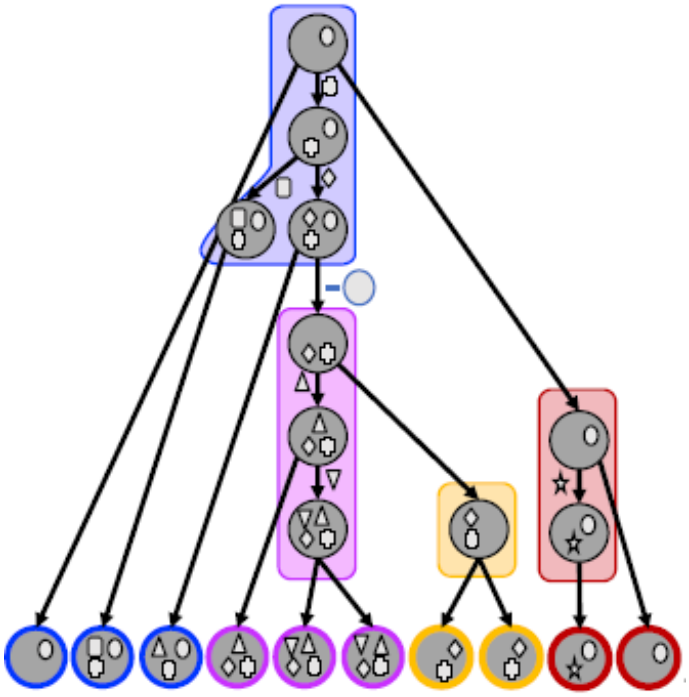


شکل ۱۵.۱: مدل احتمالاتی برای توالی‌یابی داده

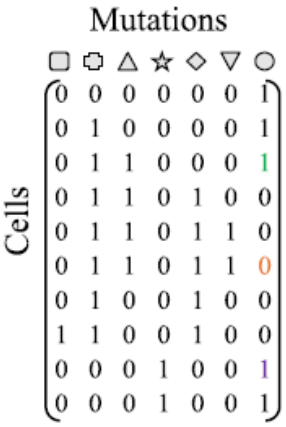
ساختن درخت اتصالات T' :

و در نهایت ماتریس جهش‌ها B* با بیشینه درست‌نمایی:

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



شکل ۱۶.۱: ساختن درخت اتصالات T'



شکل ۱۷.۱: ماتریس جهش‌ها B^* با بیشینه درست‌نمایی

$$y = xxxxxxxxxxxxxxxxxx$$

الگوریتم اسکارلت از ۲ قسمت اصلی زیر تشکیل شده است.

- محاسبه وضعیت‌های جهش با بیشینه درست‌نمایی R^* از ریشه زیردرخت‌ها[۳۳]

[33] Subtrees

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

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 جهش های معتبر برای هر موقعیت مکانی جهش a محاسبه شده و سپس بیشینه راست نمایی بالا حساب خواهد شد.

مرحله دوم:

یافتن زیردرخت‌ها پایش‌شده:

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

$y = xxxxxxxxxxxxxxxxxx$

و در نهایت حل معادله برنامه‌ریزی خطی عدد صحیح زیر:

$y = xxxxxxxxxxxxxxxxxx$

منوط به شروط زیر: $y = xxxxxxxxxxxxxxxxxx$

با فرض اینکه M عدد ثابت و بزرگی است:

$y = xxxxxxxxxxxxxxxxxx$

نقض $y = xxxxxxxxxxxxxxxxxx\ F_{w,a,b}$

نقض $G_{w,a,b}$

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

نقض $H_{w,a,b}$

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

۸.۱ الگوریتم DeepPhylo [۱]

همانطور که می‌دانیم، سرطان یک بیماری تکاملی است که با تجمع تدریجی جهش بدنی[۳۴] در سلول های تومور مشخص می‌شود. رمزگشایی از تاریخچه تکاملی یک تومور، یک چالش مهم در مطالعات سرطان است و می تواند از جنبه‌های مهم بالینی از جمله پیشرفت تومور[۳۵]، گشترش متاستاتیک[۳۶] و وجود زیرکلون‌های واگرا[۳۷] در شاخه‌های مختلف درخت فیلوژنتیک تومور درک بهتری از تومور در اختیار ما بگذارد. با توجه به اهمیت مسئله، تحولات سریعی در طراحی روش‌های محاسباتی اصولی برای استنباط فیلوژنی تومور وجود داشته است. بسیاری از این روش‌ها از داده‌های توالی‌یابی‌های انبوه[۳۸] استفاده می‌کنند که DNA میلیون‌ها سلول سرطانی و طبیعی با هم یک توالی را تشکیل می‌دهند. استنباط درخت فیلوژنی با استفاده از این نوع داده‌ها، معمولاً بر مبنای دگرگونی‌های شناسایی‌شده[۳۹] از بخش‌های مختلف سلول‌های سرطانی انجام می‌شود. به عنوان مثال: حذف و تغییر تک‌نوکلئوتیددها[۴۰] [۳۱، ۹، ۱۷، ۶، ۲۵، ۱۱]، حذف و تغییر تعداد کپی [۳۵]، دگرگونی‌های ساختاری[۴۱] [۷، ۲۱].

اگرچه استنباط درخت فیلوژنی با استفاده از این نوع داده مقرون به صرفه است اما رزولوشن[۴۲] 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 فیلوژنی تومور درنظر می‌گیرد. [۶]

با این حال، دانستن اینکه آیا تومور شامل زیرکلون‌های واگرایی است که از طریق شاخه‌های متمایزی از فیلوژنی 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

[34] Somatic mutation
[35] Tumor progression
[36] Metastatic spread
[37] Divergent subclones
[38] Bulk sequencing data
[39] Detected variants
[40] Single nucleotide variants (SNV)
[41] Structural variant
[42] Resolution

مشکل کلی استنباط کامل تاریخ تکامل تومور را فراهم می‌کند، حتی زمانی که تمام سلول‌های تک توالی که از یک نمونه بیوپستی[۴۳] توموری استخراج شده باشد. روش‌های متعددی برای استنباط تاریحچه 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 lrSCITE [۱۲] و الگوریتم OncoNEM [۲۲].
- رویکردهایی که از فرض مکان‌های بی‌نهایت استفاده نمی‌کنند و فرض را بر این می‌گذارند که تخطی‌های در شکل‌گیری درخت تکاملی فیلوژنی تا یک مقدار خطا مشخص وجود دارد، مثل الگوریتم SiFit [۳۷].

به تازگی الگوریتم‌هایی مثل SPhyR که از یک رویکرد بهینه‌سازی ترکیبی مبتنی بر زوجیت دولو[۴۴] استفاده می‌کنند یا الگوریتم SiCloneFit که بهینه یافته الگوریتم SiFit می‌باشد، ارائه شده است. [۸، ۳۶]

شایان ذکر است که روش‌های همچون PhISCS-BnB ، که از روش‌های بهینه‌سازی بر مبنای شاخه-مرز[۴۵] استفاده می‌کنند، و یا روش‌هایی مثل ScisTree، که بر مبنای اتصال اکتشافی همسایگی[۴۶] عمل می‌کند، به منظور بهبود زمان محاسباتی استنباط درخت فیلوژنی تومور ارائه شده‌اند. [۲۳، ۳۳]

در حالتی که هم داده‌های توالی‌یابی‌های انبوه و هم داده‌های توالی‌یابی تک سلولی موجود باشد می‌توان تقریب دقیق‌تری از درخت فیلوژنی تومور بدست آورد. [۱۵، ۱۸]

همانطور که در بالا خلاصه شد، روش‌های موجود برای بازسازی فیلوژنی تومور با استفاده از داده‌های توالی‌یابی تک سلولی محدودیت‌های مهمی دارند. اولاً ، بسیاری از این روش‌ها، فرض مکان‌های بی‌نهایت را به کار می‌گیرند (حتی در مواقعی که شرایطی برای خطای محدود[۴۷] و افزایش همزمان جهش‌ها[۴۸] در نظر گرفته شود) و سطح نویز یکنواختی را در نظر می‌گیرند (منفی کاذب و همچنین نرخ مثبت کاذب) هر دو این محدودیت ها، با پیشرفت درک ما از تکامل تومور و فناوری توالی‌یابی تک سلولی تغییر می‌کند. مهمتر از همه ، هدف از این روش‌ها استنباط محتمل‌ترین درخت فیلوژنی توموری است و برای حذف نویز (به دلیل مثال ، ترک آلل یا پوشش توالی کم[۴۹]) از روش‌های همچون بیشینه درست‌نمایی یا حداکثر زوجیت[۵۰] استفاده می‌کنند. به بیان

[43] Biopsy
[44] Dollo parsimony
[45] Branch-bound
[46] Joining–based heuristic
[47] Limited loss
[48] concordant gain of mutations
[49] Low sequence coverage
[50] Maximum parsimony

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 فیلوژنی تومور، فقط ویژگی‌های اساسی توپولوژی فیلوژنی تومور را استنباط کنیم، این روش‌ها نمی‌توانند به راحتی داده‌های توالی‌یابی تک سلولی شامل چند صد جهش و سلول را کنترل کنند. در نتیجه ، تکنیک‌های سریع برای استنباط ویژگی‌های کلیدی فیلوژنی 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 فیلوژنی کامل لازم است یا خیر. سرانجام، هر یک از ابزارهای موجود به تلاش انسانی 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[51] تعمیم قابل توجهی از فرمول‌بندی‌ها را برای حل بسیاری از مشکلات نشان داده است. [۲۹، ۵، ۱۴]

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 فیلوژنی های 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 [۳] و پردازش زبان طبیعی. [۵]

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 بشترین استفاده را در جهت حل مسائل خواهند کرد. [۲۸]

[51] Deep learning

با این حال این مقاله، اولین مقاله استنباط درخت فیلوژنی تومور مبتنی بر رویکردهای داده محور است. در این مقاله، اولین روش‌های بازسازی فیلوژنی تومور مبتنی بر داده را برای رفع محدودیت‌های استراتژی‌های موجود ارائه شده است. نویسندگان این مقاله از داده‌های توالی‌یابی تک سلولی در کنار شبکه‌های عصبی عمیق و یادگیری تقویتی برای استنباط ویژگی‌های توپولوژیکی فیلوژنی 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 شبیه‌سازی شده داده‌های توالی‌یابی تک سلولی وجود دارد.

۳. نویز و خطاهای موجود در داده های توالی یابی 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، می توان نظارت را به صورت مجموعه داده‌ای از ورودی‌های نویزدار به همراه با ورودی‌های بدون نویز ارائه داد. یک نظارت جایگزین و ارزان تر توسط مکانیزم بازخورد[۵۲] است که تعیین می‌کند که آیا یک خروجی از شبکه عصبی با موفقیت بدون نویز شده است یا خیر. گزینه سوم توسط یک تابع هزینه ارائه می‌شود که به طور غیرمستقیم 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[۳۲]، رویکرد NeuroSAT [۲۷] برای مساله رضایت‌مندی با

[52] feedback

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[۵۳] به کار گرفته شد. تابع هزینه استفاده شده در این مدل یک تابع هزینه خاص برای رفع مساله از بین بردن نویز بود.

$$y = xxxxxxxxxxxxxxxxxxx$$

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

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.

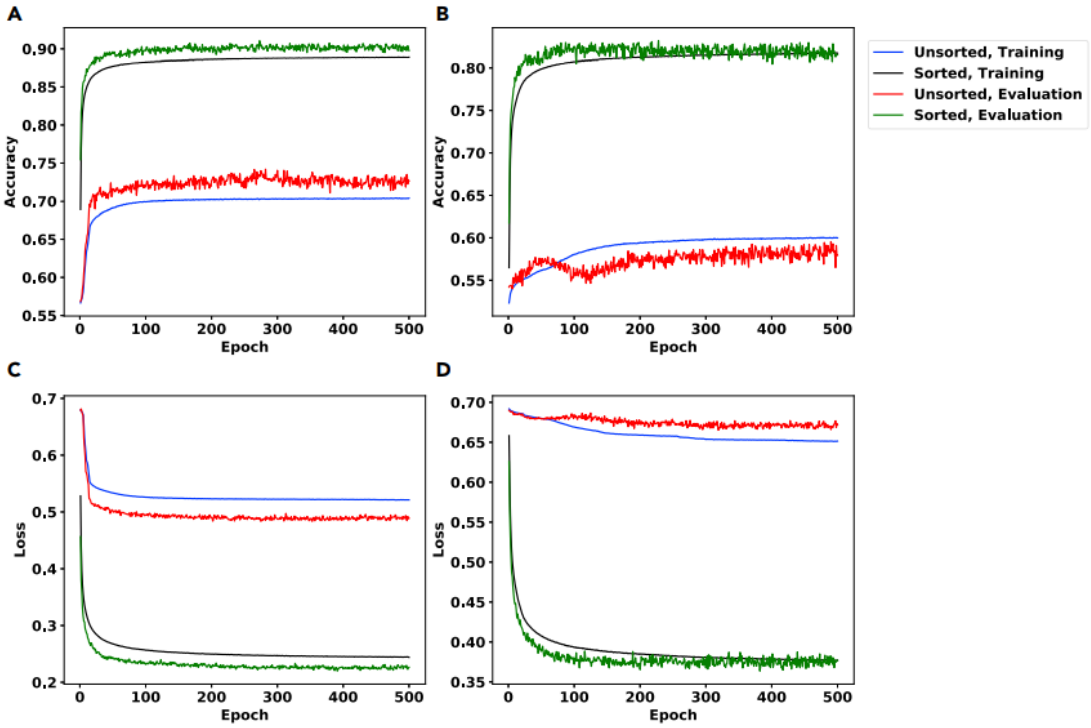
۴. داده‌های شبیه‌سازی شده ورودی مدل از طریق یک چاچوربی که راستی‌آزمایی شده است، توسعه یافته است.

در نمودار ۱۸.۱، 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 داده‌هایی را نشان می‌دهد که با نزخ نویزهای $\alpha = 0.02$ و $\beta = 0.1$ نمونه‌برداری شده‌اند اما تصاویر B و C میزان دقت شبکه در حذف نویز داده‌هایی را نشان می‌دهد که با نرخ‌های کاذب مثبت $\alpha = 0.00004$ و کاذب منفی $\beta = 0.002$ نمونه‌برداری شده است.

همچنین در جدول ۲.۱ 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.

در نهایت مقایسه بین عملکرد الگوریتم پیشنهادی در این مقاله و الگوریتم PhISCS با استفاده از معیار شباهت MLTSM84 انجام شد که نتجیه این مقایسه در شکل ۱۹.۱ آمده است. همانطور که در شکل ۱۹.۱

[53] Gold standard



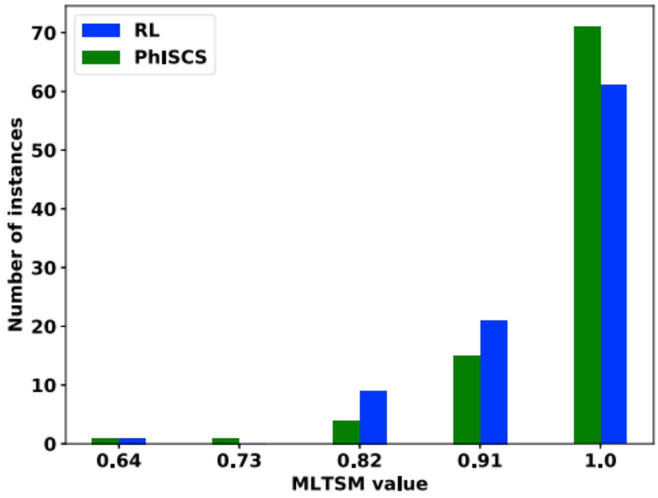
شکل ۱۸.۱: عع

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

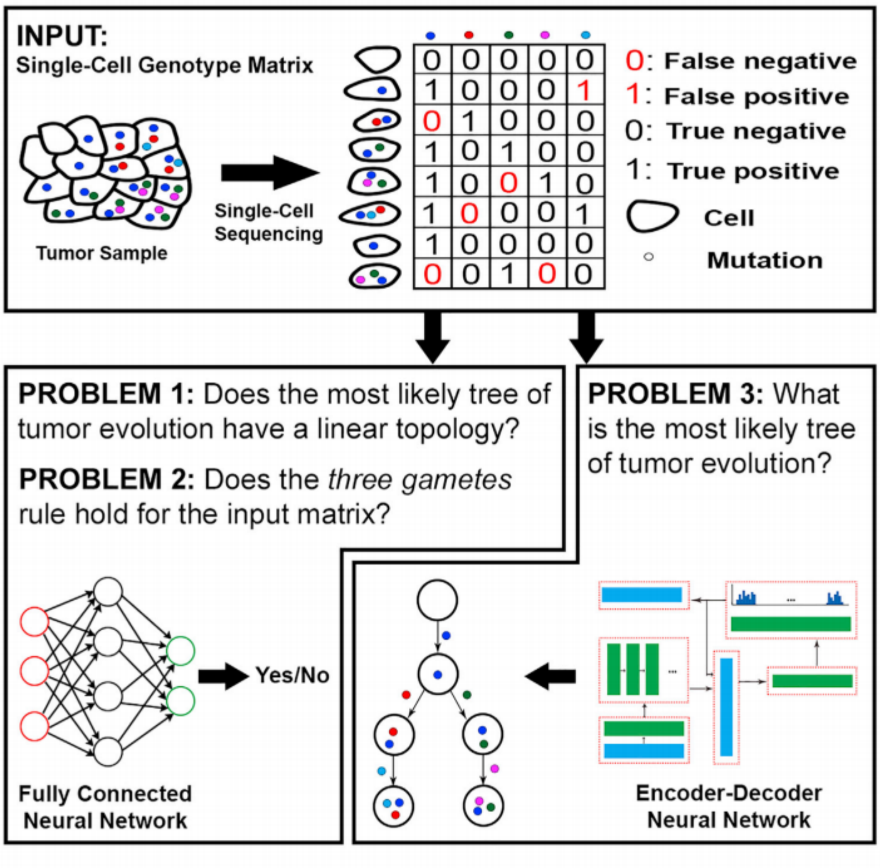
ffffffffffffffffffffffff 1.2:

مشهود است عملکرد الگوریتم پیشنهادی در میزان شباهت‌های مشابه، تعداد استنباط‌های بیشتری از فیلوژنی تومور را شامل می‌شود.

خلاصه‌ای از سازکار به کار رفته در این مقاله در شکل ۲۰.۱ آمده است:



شکل ۱۹.۱: ععع



شکل ۲۰.۱: عع

۹.۱ 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 فیلوژنی توموری از داده های توالی‌یابی تک سلولی DNA فقط از دگرگونی‌های تک‌هسته‌ای استفاده می‌کردند. [۳۰، ۱۶، ۲۰، ۸] و جهش‌های حذف و تغییر تعداد کپی و در نتیجه اطلاعات مهم استنباط فیلوژنیک 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 زیرنواحی را تحت تاثیر قرار نمی‌دهند. با مطالعه بر روی تومورهای مختلف این امکان حاصل می‌شود تا الگوهای درون توموری بهتر شناخته شوند و درمان دارویی تومور کارآمدتر و بهینه‌تر از گذشته صورت بپذیرد. مطالعه بر روی داده‌های توالی‌یابی تک سلولی 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 تک سلولی 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 ۱.۳: Comparison

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

فصل ۲

روش پیشنهادی

۱.۲ 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 بپردازیم. 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 می‌پردازد.

۲.۲ 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[1] هستند که حاوی اطلاعات جهش‌های ماتریس

[1] Ascii

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

در ادامه جدول ۱.۲ را برای معرفی اندیس‌های بکار گرفته شده در روابط مربوط به روش پیشنهادی اول معرفی می‌نماییم.

جدول ۱.۲: اندیس‌های به کار رفته در روابط روش پیشنهادی اول

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

۳.۲ 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 مددریت 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 معرفی روشی برای بدست آوردن درخت فیلوژنی 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 روشی پیشنهادی بررسی خواهد شد.

۱.۳.۲ 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 لزوما 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 خطا 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 خطا 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 جایگزاری اشتباهی انجام 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 درنظر 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 $b(i) = ۰$ باشد، در این صورت $a(i)$ حتما باید ۰ باشد وگرنه فرض اولیه مدل مکان‌های بی‌نهایت نقض خواهد شد. اما اگر $b(i) = ۱$ باشد، آنگاه نتیجه خاصی نمی‌توان گرفت و باید به دنبال نمونه والد a یعنی نمونه d باشیم. حال اگر $d(i) = ۱$ باشد، آنگاه $a(i)$ حتما باید ۱ باشد. اما اگر $d(i) = ۰$ بود آنگاه انتخاب هر مقداری برای $a(i)$ تقریبا آزاد خواهد بود زیرا با فرض اولیه تناقضی ندارد و اینکه ساختار فیلوژنی را تغییر نمی‌دهد. اما از آنجایی که خود داده‌های در دسترس شامل خطا می‌باشند و هر نمونه‌ای که حاوی اطلاعات از دست رفته است لزوما یک نواده یا یک والد ندارد، مجموعه‌ای از سلول‌های فرزند

یا والد خواهند بود که متناسب با پارمترهای خطایی که در نظر می‌گیریم و فاصله ژنی‌ای 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 i برای سلول j نظر دهد.

$$F_s(D_{ij}) = \sum_{n \in \mathcal{N}} (1 - D_{mj}) \prod_{m=1}^{M} W(D_{mn}, D_{mj}) \tag{۱.۲}$$

در فرمول ۱.۲ 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 ۲.۲ تعریف شده است.

$$W(c,p) = \begin{cases} 1 & \text{if} \quad c = 1, p = 1 \\ 1 - \xi & \text{if} \quad c = 1, p = 0 \\ 0 & \text{if} \quad c = 0, p = 1 \\ 1 & \text{if} \quad c = 0, p = 0 \end{cases} \tag{۲.۲}$$

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

به همین صورت برای اولاد سلول j نیز می‌توان مشابه حالت قبل عمل کرد که روابط آن به صورت فرمول ۳.۲

خواهد شد.

$$F_a(D_{ij}) = \sum_{n \in \mathcal{N}} D_{mj} \prod_{m=1}^{M} W(D_{mj}, D_{mn}) \tag{۳.۲}$$

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 آن‌جایگاه ژنی از محاسبه آن خود داری شود و رویکرد دوم استفاده از از مقدار ۰٫۵ یا فراوانی نسبی آن جهش در محسبات است که ما رویکرد اول را در این گزارش استفاده خواهیم کرد.

نکته دوم وجود خطا در داده‌هاست. برای مدیریت این مشکل می‌توان با مدل‌سازی خطا که به صورت فرمول ۴.۲ بیان می‌شود، برخورد کرد.

$$\begin{aligned} P(D_{ij} = 1 | E_{ij} = 0) &= \alpha, & P(D_{ij} = 0 | E_{ij} = 0) &= 1 - \alpha \\ P(D_{ij} = 0 | E_{ij} = 1) &= \beta, & P(D_{ij} = 1 | E_{ij} = 1) &= 1 - \beta \end{aligned} \tag{۴.۲}$$

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

$$W_e(c, p) = \sum_{i,j \in \{0,1\}} P(c | E_c = i) P(p | E_p = j) W(i, j) \tag{۵.۲}$$

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

$$\begin{aligned} \hat{F}_s(D_{ij}) &= \sum_{n \in \mathcal{N}} [1 - D_{mj}(1 - \alpha)] \prod_{m=1}^{M} W_e(D_{mn}, D_{mj}) \\ \hat{F}_a(D_{ij}) &= \sum_{n \in \mathcal{N}} D_{mj}(1 - \beta) \prod_{m=1}^{M} W_e(D_{mj}, D_{mn}) \end{aligned} \tag{۶.۲}$$

حال پس از محاسبه مقادیر $\hat{F}_s$ و $\hat{F}_a$ می‌توان در مورد داده نامعلوم D_{ij} به صورت فرمول ۷.۲ تصمیم گرفت.

$$D_{ij} = \begin{cases} ۰ & \text{if} \quad \hat{F}_s \geq \hat{F}_a \\ ۱ & \text{if} \quad \hat{F}_s < \hat{F}_a \end{cases} \tag{۷.۲}$$

همچنین با کمی دقت در فرمول‌بندی انجام شده اگر برای تمام i, jهای 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.

۲.۳.۲ تصادفی

پر کردن کاملا 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 مدل‌سازی‌ای که از قبل در نظر گرفته بودیم با این روش نادقیق شوند.

۴.۲ 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[۲] پشت سر هم هستیم. برای مشخص شدن نحوه کارکرد روش پیشنهادی در مراحلی که در ادامه بیان

[2] Iteration

خواهد شد به عنوان مثال یک ماتریس

$$D = \begin{bmatrix} ۰ & ۱ & ۱ & ۰ \\ ۰ & ۰ & ۱ & ۱ \\ ۱ & ۰ & ۱ & ۰ \\ ۱ & ۱ & ۰ & ۱ \\ ۱ & ۰ & ۱ & ۱ \\ ۰ & ۱ & ۱ & ۱ \end{bmatrix} \quad (۸.۲)$$

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.)

۱.۴.۲ 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 داده‌هایی که miss شده‌اند با یکی از دو روشی که معرفی شد تخمین رده می‌شوند و برای ورود به مرحله بعد آماده می‌شوند.

۲.۴.۲ 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 ابزرهایی 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 فیلوژنی هست با توجه پارامترهای ورودی و انتخاب یک نود (ژن) $root$ به عنوان ریشه این درخت به صورت زیر حاصل می‌شود.

$$\begin{aligned} &\mathcal{M} = \{۱ \dots M\} \\ &\hat{B}_{T۱} = [R_۱(\mathcal{M} - |۱|), R_۲(\mathcal{M} - |۲|), \dots, R_{root}(\{\}), \dots, R_M(\mathcal{M} - |M|)] \end{aligned} \tag{۹.۲}$$

که در این رابطه $\mathcal{M}$ برابر با مجموعه تمامی جهش‌های متمایز از شماره ۱ تا M است و $\hat{B}$ مشخص‌کننده نود پدر در درخت برای جهش iام در این لیست خود است که توسط تابع $R_i(X)$ به صورت کاملا یکنواخت[۳] از اعضای مجموعه X انتخاب می‌شود.

با توجه به مثالی که در رابطه ۸.۲ زده شد فرض کنید مقدار بردار $\hat{B}$ با ریشه $root = ۲$ به صورت رابطه ۱۰.۲ شود.

$$\hat{B} = [۲, ۱, -, ۰] \tag{۱۰.۲}$$

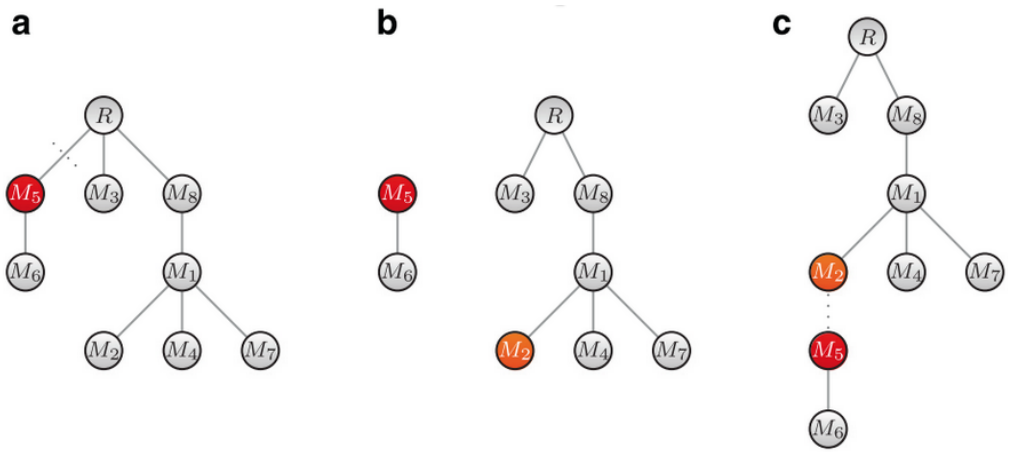
که درخت شکل ۱.۲ را نتیجه می‌دهد.

شکل ۱.۲: درخت تصادفی اول

۳.۴.۲ پاسخی جدید

تا به اینجا ما یک درخت فیلوژنی 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 تقریبا مشابه با روش [۴] به صورت هرس و اتصال دوباره[۴] قصد داریم تا درخت پاسخ فعلی را برای رسیدن به یک پاسخ دیگر تغییر دهیم. در شکل ۲.۲ 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

[3] Uniform
[4] Prune and reattach



شکل ۲.۲: نحوه انجام کار روش هرس و اتصال دوباره [۴].

پدرش قطع می‌شود. در این حالت به دو درخت مشابه با شکل b می‌رسیم. حال در درخت باقی مانده یک نود دیگر (M_2) به عنوان پدری جدید انتخاب می‌شود تا با این تغییر به درخت جدید شکل c برسیم.

در نتیجه با این تکنیک می‌توان به پاسخ‌های جدید رسید اما سوالی که باقی ماند این است که این دو نود چگونه باید انتخاب شوند؟ این دو انتخاب به صورت هوشمند توسط دو شبکه عمیق گرفته می‌شود. این شبکه‌ها که در اصل شبکه‌های یادگیری تقویتی[۵] 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 ××××××× دو درخت راهبر-پیرو[۶] مکان‌های مناسب برای بازاتصال را ارزش‌گزاری می‌کند. این دو شبکه در قسمت‌های ۷.۴.۲ و ۸.۴.۲ به تفضیل شرح داده شده‌اند.

۴.۴.۲ 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 دو زیربخشی که در ادامه آمده است بیان شده‌اند.

[5] Reinforcement learning
[6] Master-slave

۱.۴.۴.۲ 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 متشکل از مقادیر ۰ و ۱ خواهد بود که به عنوان مثال $X_{i,j} = ۱$ 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 $\wp$ مزین کنیم و به ماتریس‌های مختلفی برسیم. 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 $\wp$ را بدست آوریم و از روی آن‌ها ماتریس متناظر X را بدست آوریم. از پیش با بررسی پروفایل‌های شماره کپی[۷] در بخش ۵.۴.۲ 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 ۱۱.۲ تعریف می‌کنیم که برای انتخاب بهینه σ مورد استفاده قرار خواهد گرفت.

$$B_{i,j} = \begin{cases} ۱ & \text{if} \quad i = j \text{ یا } i \text{ یکی از نوادگان } j \text{ باشد که در } \mathcal{L} \text{ نباشد} \\ x & \text{if} \quad i \text{ یکی از نوادگان } j \text{ باشد و همچنین در } \mathcal{L} \text{ باشد} \\ ۰ & \text{if} \quad \text{در صورتی که دو مورد بالایی نباشد} \end{cases} \tag{۱۱.۲}$$

در رابطه بیان شده x به معنای این است که هم می‌تواند مقدار ۰ 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 ۱۲.۲ بدست می‌آید.

$$S(c_i, T, k) = \prod_{j=۰}^{M} P(D_{i,j}|B_{j,k}) \tag{۱۲.۲}$$

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

$$P(D = ۱|B = x) = ۱ - \beta, \qquad P(D = ۰|B = x) = ۱ - \alpha \tag{۱۳.۲}$$

[7] Copy 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 $\wp$ 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 ۱۱.۲ را به صورت رابطه ۱۴.۲ بازنویسی می‌کنیم.

$$B_{i,j} = \begin{cases} ۱ & \text{if} \quad \text{اگر } i \text{ یکی از نوادگان } j \text{ باشد که در } \mathcal{L} \text{ نباشد} \\ x_i^{\text{dist}(i,j)} & \text{if} \quad \text{اگر } i \text{ یکی از نوادگان } j \text{ باشد و همچنین در } \mathcal{L} \text{ باشد} \\ ۰ & \text{if} \quad \text{در صورتی که دو مورد قبلی نباشد} \end{cases} \tag{۱۴.۲}$$

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

$$A_{i,j} = \begin{cases} ۱ & \text{if} \quad \wp_i < \text{dist}(i,j) \text{ یا } B_{i,j} = ۱ \\ ۰ & \text{if} \quad \text{شرط بالا درست نباشد} \end{cases} \tag{۱۵.۲}$$

این مقادیر نامشخص که در اتصال به ژن i و نوادگان آن در درخت مشخص شده‌اند با توجه به فاصله تعیین شده از این ژن i برای نوادگان حذف خواهد شد که این فاصله در $\wp_i$ مشخص شده است. پس حال با تعیین مقادیر بردار $\wp$ می‌توانیم بهترین B نامعلوم را به A معلوم تبدیل کنیم. به رابطه ۱۲.۲ توجه کنید. این رابطه امتیاز اتصال نمونه 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 iامین 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 ارمغان می‌آورد مشخص کنیم

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

$$\begin{aligned} S(c_i, T) &= \max_{k \in \{0 \dots M\}} S(c_i, T, k) \\ &= \max_{j^*} \left(\prod_{k=0}^{M} P(D_{i,k} | A_{k,j^*}) \right) = S(c_i, T, \sigma_i) \end{aligned} \tag{۱۶.۲}$$

رابطه ۱۶.۲ 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 $\wp$ یافته‌ایم. آیا مقدار $\wp$ نیز بهینه است؟ برای مشخص کردن مقدار بهینه $\wp$ برای درخت دلخواه T از رابطه‌ای که در ادامه آمده است کمک می‌گیریم.

$$\langle \hat{\wp}, \hat{\sigma} \rangle = \arg\max_{\wp, \sigma} \prod_{i=1}^{N} S(c_i, T) \tag{۱۷.۲}$$

در واقع این مقادیر $\wp$ 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 $\hat{\wp}$ 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 ۱۸.۲ تشکیل داد.

$$X_{i,j} = \hat{A}_{i,\hat{\sigma}_j} \tag{۱۸.۲}$$

۵.۴.۲ یافتن جهش‌های با پتانسیل حذف

همان‌گونه که از ابتدا می‌دانیم ما به دنبال درخت فیلوژنی 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 $(\wp)$
- نحوه انتصاب نمونه‌های مشاهده شده به درخت (σ)

- و در نهایت پارامترهایی که برای مدل‌سازی خطای بوجود آمده در داده‌های در دسترس‌مان تعیین شده است (θ)

به گونه‌ای انتخاب شوند که محتمل‌ترین حالت را برای مشاهده داده‌های 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.

۱.۵.۴.۲ 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 ۱۹.۲ محاسبه می‌شود.

$$L : P(D|T, \sigma, \wp, \theta) = \prod_{n=1}^{n=N} \prod_{m=1}^{m=M} P(D_{nm}|X_{nm}) \qquad (۱۹.۲)$$

که X برابر ماتریس بدست آمده از درخت T با توجه به بردارهای σ و $\wp$ است. این رابطه بیانگر احتمال مشاهده ماتریس داده ورودی D در صورتی است که درخت فیلوژنی صحیح T و پارامترهای حقیقی θ باشد که توسط بردارهای σ و $\wp$ تثبیت 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 Tای به همراه بردارهای مربوطه آن هستیم که پاسخ رابطه باشد.

$$(T, \sigma, \wp)_{\text{ML}} = \arg \max_{(T,\sigma,\wp)} P(D|T, \sigma, \wp, \theta) \qquad (۲۰.۲)$$

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

$$S(T) = P(D|T, \sigma^*, \wp^*), \qquad \sigma^* = \arg \max_{\sigma} P(D|T, \sigma, \wp) \qquad (۲۱.۲)$$

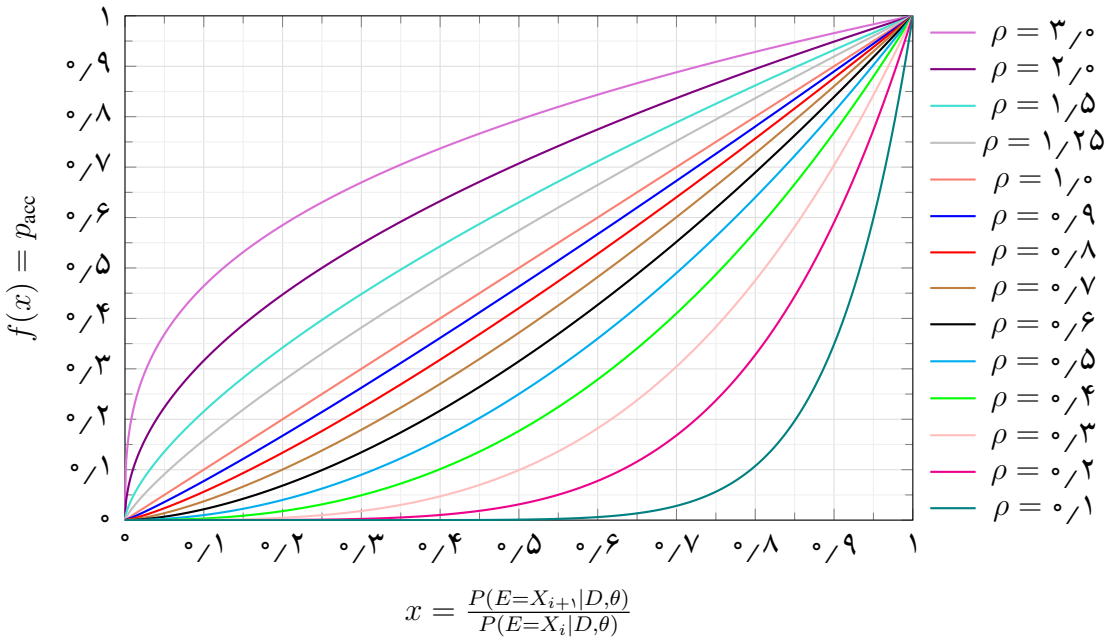
۶.۴.۲ پذیرش پاسخ‌های جدید و یافتن بهترین پاسخ

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

$$p_{\text{acc}} = \min\left[1, \left(\frac{P(E = X_{i+1}|D,\theta)}{P(E = X_i|D,\theta)}\right)^{\rho^{-1}}\right] \qquad (۲۲.۲)$$

در رابطه ۲۲.۲، 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 حتما 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 مستقیما رد نمی‌کنیم و به احتمالی کمتر از ۱ ممکن است آن را بپذیریم. دلیل این پذیرش جلوگیری از به دام افتادن الگوریتم در پاسخ‌مان در بیشینه‌های محلی[8] است. در شکل تاثیر تغییر پارامتر ρ در احتمال پذیرش پاسخ‌های جدیدی که مطلوب‌تر از پاسخ فعلی نیستند نمایش داده شده است.

[8] Local maxima



شکل ۳.۲: نمودار تغییر احتمال پذیرش پاسخ جدید نامطلوب‌تر با توجه به مقدار پارامتر ρ.

۷.۴.۲ شبکه هرس‌کننده

در روش‌های گذشته مانند روش‌های ارائه شده در [۳۵] و [۱۲] 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 جست‌وجو 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.

۱.۷.۴.۲ ورودی

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

(۲۳.۲)

$$\begin{bmatrix} I_{۱,۱} & I_{۱,۲} & \dots & I_{۱,j} & \dots & I_{۱,M} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ I_{i,۱} & I_{i,۲} & \dots & I_{i,j} & \dots & I_{i,M} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ I_{N,۱} & I_{N,۲} & \dots & I_{N,j} & \dots & I_{N,M} \end{bmatrix}_{N \times M} \rightarrow \begin{bmatrix} ۱ & ۱ & f(I_{۱,۱}) & f(I_{۱,۱}) \\ ۱ & ۲ & f(I_{۱,۲}) & f(I_{۱,۲}) \\ \vdots & \vdots & \vdots & \vdots \\ i & j & f(I_{i,j}) & f(I_{i,j}) \\ \vdots & \vdots & \vdots & \vdots \\ N & M & f(I_{N,M}) & f(I_{N,M}) \end{bmatrix}_{N*M \times ۱}$$

در صورتی که

$$f(I_{i,j}) = \begin{cases} \alpha & \text{if } X_{i,j} - D_{i,j} = -۱ \text{ (false positive)} \\ \alpha & \text{if } X_{i,j} + D_{i,j} = ۰ \text{ (true negetive)} \\ \beta & \text{if } X_{i,j} - D_{i,j} = ۱ \text{ (false negetive)} \\ \alpha & \text{if } X_{i,j} + D_{i,j} = ۲ \text{ (true positive)} \end{cases} \quad (۲۴.۲)$$

این تبدیل در رابطه ×××× بیان شده است.

a to A transform We A . matrix binary a as given is problem this to input that Recall A of row Each network. neural the to it feeding before below، shown as A matrix new A in row given a of columns two first The A . of entries the of one exactly to corresponds third The A . in entry corresponding the of column the and row the represent respectively is which - entry the of value actual the on depends and level noise the represents column either is column third the of value the case، simplest the In column. last the in represented case، general more the In zero. value has entry the if - is or one، value has entry the if -

the of entry each for rate negative false or positive false distinct a specify can user the training for ability our to key is transformation this below، depicted As A . matrix input shapes/dimensions: varying of matrices with network neural the

۸.۴.۲ شبکه بازاتصال‌کننده

۹.۴.۲ 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 گام های 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 خطا 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 خطا در روش آن وجود داشت اما ما برای بهتر کردن امتیاز به جای تغییر پارامترهای خطا به ازای یک درخت جست و جوی ضمیمه کردن های مختلف سلول‌ها و لاس شدن جهش ها 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 فیلوژنی 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 فیلوژنی 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 مدل های دولو یا فرض مکان‌های بی‌نهایت، ابهام کمتری در استنباط درخت فیلوژنی 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 فیلوژنی تومور منجر به افزایش دقت فیلوژنی استنباط شده است. در این مقاله نشان داده شد که فیلوژنی توموری استنباط شده برای بیماران مبتلا به سرطان روده از دقت و تکرارپذیری بیشتری برخوردار است و این الگوریتم در نهایت فیلوژنی‌هایی 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 مواقعی 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.

مراجع

[1] Azer, Erfan Sadeqi, Ebrahimabadi, Mohammad Haghir, Malikić, Salem, Khardon, Roni, and Sahinalp, S Cenk. Tumor phylogeny topology inference via deep learning. *Iscience*, 23(11):101655, 2020.

[2] Beerenwinkel, Niko, Schwarz, Roland F, Gerstung, Moritz, and Markowetz, Florian. Cancer evolution: mathematical models and computational inference. *Systematic biology*, 64(1):e1–e25, 2015.

[3] 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.

[4] Davis, Alexander and Navin, Nicholas E. Computing tumor trees from single cells. *Genome biology*, 17(1):1–4, 2016.

[5] 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.

[6] 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.

[7] Eaton, Jesse, Wang, Jingyi, and Schwartz, Russell. Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics*, 34(13):i357–i365, 2018.

[8] El-Kebir, Mohammed. Sphyr: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics*, 34(17):i671–i679, 2018.

[9] 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.

[10] 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.

[11] 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.

[12] Jahn, Katharina, Kuipers, Jack, and Beerenwinkel, Niko. Tree inference for single-cell data. *Genome biology*, 17(1):1–17, 2016.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] Malikic, Salem, Mehrabadi, Farid Rashidi, Azer, Erfan Sadeqi, Ebrahimabadi, Mohammad Haghir, and Sahinalp, S Cenk. Studying the history of tumor evolution from single-cell sequencing data by exploring the space of binary matrices. *bioRxiv*, 2020.

[18] 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.

[19] McGranahan, Nicholas and Swanton, Charles. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*, 168(4):613–628, 2017.

[20] 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.

[21] 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.

[22] Ross, Edith M and Markowetz, Florian. Onconem: inferring tumor evolution from single-cell sequencing data. *Genome biology*, 17(1):1–14, 2016.

[23] 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.

[24] 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.

[25] Satas, Gryte and Raphael, Benjamin J. Tumor phylogeny inference using tree-constrained importance sampling. *Bioinformatics*, 33(14):i152–i160, 2017.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] Singer, Jochen, Kuipers, Jack, Jahn, Katharina, and Beerenwinkel, Niko. Single-cell mutation identification via phylogenetic inference. *Nature communications*, 9(1):1–8, 2018.

[31] 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.

[32] Williams, Ronald J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[33] 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.

[34] 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.

[35] 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.

[36] Zafar, Hamim, Navin, Nicholas, Chen, Ken, and Nakhleh, Luay. Siclonefit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome research*, 29(11):1847–1859, 2019.

[37] 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.