

دانشگاه تهران  
پردیس دانشکده‌های فنی  
دانشکده علوم و فنون نوین  
گروه شبکه



## استنتاج درخت فیلوزنی تومور سرطانی با استفاده از داده‌های تکسلولی و تغییرات تعداد تکرار

پایان‌نامه برای دریافت درجه کارشناسی ارشد در رشته مهندسی فناوری اطلاعات  
گرایش سامانه‌های شبکه‌ای

افشین بزرگپور

اساتید راهنما

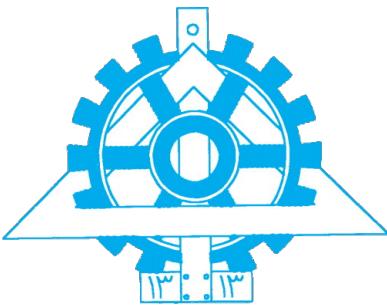
دکتر سامان هراتی‌زاده و دکتر ابوالفضل مطهری

۱۴۰۰ مرداد



سُبْحَانَ رَبِّ الْجَمَلِ





دانشگاه تهران  
پردیس دانشکده‌های فنی  
دانشکده علوم و فنون نوین  
گروه شبکه



## استنتاج درخت فیلوزنی تومور سرطانی با استفاده از داده‌ای تک‌سلولی و تغییرات تعداد تکرار

پایان‌نامه برای دریافت درجهٔ کارشناسی ارشد در رشته مهندسی فناوری اطلاعات  
گرایش سامانه‌های شبکه‌ای

افشین بزرگ‌پور

اساتید راهنما

دکتر سامان هراتی‌زاده و دکتر ابوالفضل مطهری

۱۴۰۰ مرداد





دانشگاه تهران  
پردیس دانشکده‌های فنی  
دانشکده علوم و فنون نوین

## گواهی دفاع از پایان‌نامه کارشناسی ارشد

هیأت داوران پایان نامه کارشناسی ارشد آفای / خانم افشین بزرگ پور به شماره دانشجویی ۸۳۰۵۹۶۰۰۵ در رشته مهندسی فناوری اطلاعات - گرایش سامانه های شبکه ای را در تاریخ ..... با عنوان «استنتاج درخت فیلوژنی تومور سرطانی با استفاده از داده های تک سلولی و تغییرات تعداد تکرار»

با نمره نهایی	به عدد	به حروف

و درجه ارزیابی کرد.

ردیف	مشخصات هیأت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضا
۱	استاد راهنمای	دکتر سامان هراتی‌زاده	استادیار	دانشگاه تهران	
۲	استاد راهنمای	دکتر ابوالفضل مطهری	استادیار	دانشگاه تهران	
۳	استاد داور داخلی	دکتر داور داخلی	دانشیار	دانشگاه تهران	
۴	استاد مدعو	دکتر داور خارجی	دانشیار	دانشگاه داور خارجی	
۵	نماینده تحصیلات تکمیلی، دانشکده	دکتر نماینده	دانشیار	دانشگاه تهران	

## نام و نام خانوادگی معاون آموزشی و تحصیلات

## تکمیلی پردازش دانشکده‌های فنی:

تاریخ و امضا:

## نام و نام خانوادگی معاون تحصیلات تکمیلی و

پژوهشی دانشکده / گروه:

تاریخ و امضا:



## تعهدنامه اصالت اثر

باسم‌هه تعالی

اینجانب افشنین بزرگ‌پور تأیید می‌کنم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب است و به دستاوردهای پژوهشی دیگران که در این نوشه از آن‌ها استفاده شده است مطابق مقررات ارجاع گردیده است. این پایان‌نامه قبل‌اً برای احراز هیچ مدرک هم‌سطح یا بالاتری ارائه نشده است.

نام و نام خانوادگی دانشجو: افشنین بزرگ‌پور  
تاریخ و امضاء دانشجو:



کلیه حقوق مادی و معنوی این اثر  
متعلق به دانشگاه تهران است.



تقدیم به:

## همسر و فرزندانم

و

پدر و مادرم



## قدردانی

سپاس خداوندگار حکیم را که با لطف بی کران خود، آدمی را به زیور عقل آراست.  
در آغاز وظیفه خود می دام از زحمات بی دریغ اساتید راهنمای خود، جناب آقای دکتر ... و ...، صمیمانه  
تشکر و قدردانی کنم که در طول انجام این پایان نامه با نهایت صبوری همواره راهنمای و مشوق من بودند و قطعاً  
بدون راهنمایی های ارزنده ایشان، این مجموعه به انجام نمی رسید.  
از جناب آقای دکتر ... که زحمت مشاوره، بازبینی و تصحیح این پایان نامه را تقبل فرمودند کمال امتنان را  
دارم.

با سپاس بی دریغ خدمت دوستان گران مایه ام، خانم ها ... و آقایان ... در آزمایشگاه ...، که با همفکری مرا  
صمیمانه و مشفقانه یاری داده اند.

و در پایان، بوسه می زنم بر دوستان خداوندگاران مهر و مهربانی، پدر و مادر عزیزم و بعد از خدا، ستایش می کنم  
وجود مقدس شان را و تشکر می کنم از خانواده عزیزم به پاس عاطفه سرشار و گرمای امیدبخش وجودشان، که  
بهترین پشتیبان من بودند.

افشین بزرگ پور

۱۴۰۰ مرداد



## چکیده

این راهنمای نمونه‌ای از قالب پژوهش، پایان‌نامه و رساله دانشگاه تهران می‌باشد که با استفاده از کلاس-*tehran* و بسته زی پرشین در *LATeX* تهیه شده است. این قالب به گونه‌ای طراحی شده است که مطابق با دستورالعمل نگارش و تدوین پایان‌نامه کارشناسی ارشد و دکتری، مورخ ۹۳/۰۶/۰۳ پردازی دانشکده‌های فنی دانشگاه تهران باشد و حروف چینی بسیاری از قسمت‌های آن، مطابق با استاندارد قالب‌های فارسی پایان‌نامه در لاتک، به طور خودکار انجام می‌شود.

چکیده بخشی از پایان‌نامه است که خواننده را به مطالعه آن علاقمند می‌کند و یا از آن می‌گریزاند. چکیده باید ترجیحاً در یک صفحه باشد. در نگارش چکیده نکات زیر باید رعایت شود. متن چکیده باید مزین به کلمه‌ها و عبارات سلیس، آشنا، بامعنی و روشن باشد. بگونه‌ای که با حدود ۳۰۰ تا ۵۰۰ کلمه بتواند خواننده را به خواندن پایان‌نامه راغب نماید. چکیده، جدای از پایان‌نامه باید به تنها یی گویا و مستقل باشد. در چکیده باید از ذکر منابع، اشاره به جداول و نمودارها اجتناب شود. تمیز بودن مطلب، نداشتن غلط‌های املایی یا دستور زبانی و رعایت دقت و تسلسل روند نگارش چکیده از نکات مهم دیگری است که باید درنظر گرفته شود. در چکیده پایان‌نامه باید از درج مشخصات مربوط به پایان‌نامه خودداری شود. چکیده باید منعکس‌کننده اصل موضوع باشد. در چکیده باید اهداف تحقیق مورد توجه قرار گیرد. تأکید روی اطلاعات تازه (یافته‌ها) و اصطلاحات جدید یا نظریه‌ها، فرضیه‌ها، نتایج و پیشنهادها متمرکز شود. اگر در پایان‌نامه روش نوینی برای اولین بار ارائه می‌شود و تا به حال معمول نبوده است، با جزئیات بیشتری ذکر شود. شایان ذکر است چکیده فارسی و انگلیسی باید حتماً به تأیید استاد راهنما رسیده باشد.

کلمات کلیدی در انتهای چکیده فارسی و انگلیسی آورده می‌شود. محتوای چکیده‌ها بر اساس موضوع و گرایش تحقیق طبقه‌بندی می‌شود و به همین جهت وجود کلمات شاخص و کلیدی، مراکز اطلاعاتی را در طبقه‌بندی دقیق و سریع پایان‌نامه یاری می‌دهد. کلمات کلیدی، راهنمای نکات مهم موجود در پایان‌نامه هستند. بنابراین باید در حد امکان کلمه‌ها یا عباراتی انتخاب شود که ماهیت، محتوا و گرایش کار را به وضوح روشن نماید.

واژگان کلیدی حداقل ۵ کلمه یا عبارت، متناسب با عنوان، قالب پایان‌نامه، لاتک



# فهرست مطالب

ث

فهرست تصاویر

ح

فهرست جداول

خ

فهرست الگوریتم‌ها

د

فهرست برنامه‌ها

۱

فصل ۱: مقدمه

۵

فصل ۲: مبانی تحقیق

۵

۱.۲ تنوع ژنتیکی

۸

۲.۲ تکامل تومور<sup>۱</sup>

۹

۳.۲ تکنولوژی‌های توالی‌بایی و فراوانی تغییرات آلل<sup>۲</sup>

۱۰

۴.۲ ناهمگنی ژنومی تومور

۱۳

۵.۲ بازسازی زیر کلونال

۱۵

۶.۲ تغییرات تعداد کپی

۱۷

۷.۲ جهش‌های ساده بدندی

۱۸

۸.۲ ترک آللی<sup>۳</sup>

۱۹

۹.۲ مقدمه‌ای بر مدل‌سازی احتمالی

<sup>1</sup>Tumor Evolution

<sup>2</sup>Variant allele frequency

<sup>3</sup>Allelic dropout

۲۰ . . . . .	۱.۹.۲ زنجیره مارکوف مونت کارلو <sup>۴</sup>
۲۲ . . . . .	۱۰.۲ یادگیری ماشین <sup>۵</sup> و یادگیری تقویتی <sup>۶</sup>
۳۱ . . . . .	۱۱.۲ شبکه‌های عصبی بازگشتی . . . . .
۳۲ . . . . .	۱.۱۱.۲ شبکه عصبی بازگشتی چیست؟
۳۳ . . . . .	۲.۱۱.۲ مزایای شبکه عصبی بازگشتی <sup>۷</sup>
۳۳ . . . . .	۳.۱۱.۲ معایب شبکه عصبی بازگشتی
۳۴ . . . . .	۴.۱۱.۲ کاربردهای شبکه عصبی بازگشتی . . . . .
۳۴ . . . . .	۵.۱۱.۲ انواع شبکه عصبی بازگشتی . . . . .
۳۵ . . . . .	۶.۱۱.۲ حافظه‌ی کوتاه‌مدت بلند (LSTM)
۴۳ . . . . .	۱۲.۲ یادگیری تقویتی . . . . .
۴۳ . . . . .	۱.۱۲.۲ مقدمه و پیشینه تاریخی . . . . .
۴۵	<b>فصل ۳: روش‌های پیشین</b>
۴۵ . . . . .	۱.۳ مقدمه . . . . .
۴۵ . . . . .	۲.۳ روش ساخت درخت تکاملی <sup>۸</sup> با استفاده از داده‌های توالی‌بابی تک‌سلولی <sup>۹</sup> . . . . .
۴۷	<b>فصل ۴: روش پیشنهادی</b>
۴۷ . . . . .	۱.۴ مقدمه . . . . .
۴۷ . . . . .	۲.۴ معرفی دادگان ورودی . . . . .
۴۸ . . . . .	۳.۴ روش پیشنهادی برای مدیریت داده‌های از دست رفته . . . . .
۴۸ . . . . .	۱.۳.۴ روش محاسبه استاتیک . . . . .
۵۱ . . . . .	۱.۱.۳.۴ تصادفی . . . . .
۵۲ . . . . .	۴.۴ روش پیشنهادی اول (درخت بازی) . . . . .

<sup>4</sup>Markov Chain Monte Carlo (MCMC)<sup>5</sup>Machine learning<sup>6</sup>Reinforcement learning<sup>7</sup>Recurrent Neural Network<sup>8</sup>Tumor evolutionary tree inference<sup>9</sup>Single cell sequencing

۰۲	۱.۴.۴ پیش‌پردازش
۰۳	فصل ۵: نتایج تجربی
۰۳	۱.۵ پایگاه داده‌های ورودی
۰۳	۱.۱.۵ پایگاه داده مصنوعی
۰۴	۱.۱.۱.۵ ساخت درخت تصادفی
۰۷	۲.۱.۱.۵ تبدیل درخت به ماتریس ژن-سلول
۰۹	۳.۱.۱.۵ اضافه کردن نویز به ماتریس ژن-جهش
۶۱	۲.۱.۵ پایگاه داده حقیقی
۶۲	۲.۵ روش پیشنهادی بدست آوردن درخت فیلوزنی
۶۳	۱.۲.۵ استفاده از شبکه ژن‌ها برای یافتن درخت فیلوزنی
۶۳	۱.۱.۲.۵ استفاده از یک درخت نمونه ناپنهنه و تغییر اتصالات تا رسیدن به درخت فیلوزنی بنهنه
۶۶	۳.۵ نتایج تجربی
۶۶	۱.۳.۵ نتایج بر روی پایگاه داده مصنوعی
۷۰	۲.۳.۵ نتایج بر روی داده‌های حقیقی
۷۰	۱.۲.۳.۵ نتایج بهینه‌سازی درخت ژنی
۷۱	۴.۵ گام‌های آتی
۷۱	۱.۴.۵ بھبود در ساخت درخت اولیه
۷۲	۲.۴.۵ افزایش سرعت همگرایی MCMC
۷۲	۱.۲.۴.۵ تنوع در گام‌ها با استراتژی معقول
۷۳	۲.۲.۴.۵ قرار دادن احتمال وزن‌دار به ازای هر انتخاب
۷۵	فصل ۶: بحث و نتیجه‌گیری
۷۷	مراجع
سوم	واژه‌نامه فارسی به انگلیسی

واژه‌نامه انگلیسی به فارسی

هفتم

# فهرست تصاویر

۳	دو مدل برای ناهمگونی تومور	۱.۱
۶	مارپیچ دوگانه دی ان ای	۱.۲
۷	همانندسازی دی ان ای	۲.۲
۸	جهش تک نوکلئوتیدی	۳.۲
۸	تغییرات ساختاری	۴.۲
۹	درخت فیلوژنیک تومور	۵.۲
۱۰	تشخیص تغییر بدنه تک نوکلئوتیدی از طریق خوانش هم ترازی	۶.۲
۱۵	درخت کلون تومور	۷.۲
۱۹	نمایی از تطابق ژنتیکی	۸.۲
۲۴	معماری یک شبکه عصبی کانولوشنی	۹.۲
۲۶	عملیات کانولوشن <sup>۱۰</sup> در یک شبکه عصبی کانولوشنی <sup>۱۱</sup> با کرنل <sup>۱۲</sup> $5 \times 5$	۱۰.۲
۲۷	(a) تابع فعالیت ReLU <sup>۱۳</sup> و (b) تابع فعالیت سیگموید <sup>۱۴</sup>	۱۱.۲
۲۸	تابع max-pooling بر روی آرایه دو بعدی کوچک $2 \times 2$ و $m = s = 2$	۱۲.۲
۲۹	لایه حذف تصادفی <sup>۱۵</sup> با $\sigma = 0.5$	۱۳.۲
۳۲	یک نمونه باز شده شبکه عصبی بازگشتی	۱۴.۲
۳۵	ساختار شبکه عصبی بازگشتی یک به یک	۱۵.۲

<sup>10</sup>Convolution

<sup>11</sup>Convolutional neural network

<sup>12</sup>Kernel

<sup>13</sup>Activation function

<sup>14</sup>Sigmoid

<sup>15</sup>Dropout

۱۶.۲ ساختار شبکه عصبی بازگشتی یک به چند . . . . .	۳۵
۱۷.۲ ساختار شبکه عصبی بازگشتی چند به یک . . . . .	۳۶
۱۸.۲ ساختار شبکه عصبی بازگشتی چند به چند . . . . .	۳۶
۱۹.۲ ساختار LSTM . . . . .	۳۷
۲۰.۲ مازول‌های تکرار شونده در شبکه‌های عصبی بازگشتی استاندارد فقط دارای یک لایه هستند. . . . .	۳۸
۲۱.۲ مازول‌های تکرار شونده در LSTM‌ها دارای ۴ لایه هستند که با هم در تعامل می‌باشند. . . . .	۳۹
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. . . . .	۴۰
۲۳.۲ سلول حالت در مازول LSTM . . . . .	۴۰
۲۴.۲ نمایی از نحوه تاثیر ورود اطلاعات به سلول حالت . . . . .	۴۰
۲۵.۲ قدم اول در پاک کردن اطلاعات از سلول حالت در وضعیت ورودی . . . . .	۴۱
۲۶.۲ قدم دوم در اضافه کردن اطلاعات جدید به سلول حالت . . . . .	۴۲
۲۷.۲ بهروزرسانی اطلاعات در سلول حالت . . . . .	۴۲
۲۸.۲ قدم نهایی برای تولید خروجی مازول LSTM . . . . .	۴۳
۱.۵ درخت فیلوزنی تصادفی تولید شده برای $n = 20$ و $\zeta$ ‌های مختلف . . . . .	۵۵
۲.۵ درخت جهش تصادفی با پارامترهای $N = 30, M = 20, \zeta = 1, \gamma = 0, \alpha = 15$ در مقایله SCITE Navin . . . . .	۵۸
۳.۵ ماتریس‌های ژن-سلول ( $E$ ) بدست آمده از درخت‌های تصادفی ساخته شده . . . . .	۵۹
۴.۵ ماتریس‌های ژن-سلول همراه با نویز و داده‌های از دست رفته شکل ۳.۵.ب که برای ورودی مسله آماده شده است. . . . .	۶۰
۵.۵ داده‌های حقیقی SCITE Navin در مقاله SCITE Xu در مقایله . . . . .	۶۱
۶.۵ داده‌های حقیقی SCITE Xu در مقایله SCITE Navin در مقاله SCITE Xu . . . . .	۶۲
۷.۵ نمونه‌ای تصادفی از ماتریس ورودی $D$ . . . . .	۶۶
۸.۵ درخت تصادفی ایجاد شده به عنوان درخت اولیه شکل ۷.۵ . . . . .	۶۷
۹.۵ نتیجه اجرای روش پیشنهادی برای ماتریس شکل ۷.۵ . . . . .	۶۸
۱۰.۵ نتیجه اجرای روش پیشنهادی برای تعداد نمونه‌های مختلف . . . . .	۶۹

۶۹ . . . . .	۱۱.۰	نتیجه اجرای روش پیشنهادی برای تعداد نمونه‌های مختلف
۷۰ . . . . .	۱۲.۵	درخت تصادفی اولیه
۷۱ . . . . .	۱۳.۵	نمودار تغییر انرژی در طی گام‌های مختلف
۷۲ . . . . .	۱۴.۵	بهترین درخت یافته شده و خروجی الگوریتم برای مقاله SCITE
۷۴ . . . . .	۱۵.۵	درخت بدست آمده در مقاله SCITE

# فهرست جداول

۱.۴	اندیس‌های به کار رفته در روابط روش پیشنهادی اول	۵۱
۲.۴	پارامترهای مدل ریاضی	۵۲
۳.۴	متغیرهای مدل ریاضی	۵۲

# فهرست الگوریتم‌ها

خ

## فهرست برنامه‌ها



# فصل ۱

## مقدمه

تومور<sup>۱</sup> از رشد غیر طبیعی سلول با احتمال حمله یا گسترش به سایر قسمت‌های بدن تشکیل می‌شود. تومورهای بدخیم<sup>۲</sup> معمولاً سرطان<sup>۳</sup> نامیده می‌شوند. سرطان علل مختلفی از جمله تغییرات ژنتیکی، آسودگی 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<sup>۴</sup> نامیده می‌شود. ناهمگنی تومور احتمالاً برای درمان سرطان و کشف نشانگر زیستی، به ویژه در روش‌های درمانی هدفمند، تأثیراتی خواهد داشت [۲۱]. درمان‌های فعلی، سرطان را به عنوان یک بیماری همگن درمان می‌کنند [۴۳].

داروهای هدفمند در برابر زیرجمعیت‌های تک یا چند سلولی با انکوژن<sup>۵</sup> جهش‌یافته 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 [۲۱]. این زیرجمعیت‌های سلولی بدون درمان ممکن است منجر به پیشرفت تومور پس از درمان دارویی شوند [۲۱]. به عنوان مثال، رشد مجدد سلول‌های تومورزا در سرطان روده بزرگ<sup>۶</sup> سرطان پستان و گلیوبالستوم<sup>۷</sup> پس از تابش یا درمان سیکلوفسفامید مشاهده شده است [۴۳]. بنابراین، مطالعه روند رشد تومور و ناهمگنی آن تأثیرات زیادی بر تشخیص و درمان سرطان دارد.

تومورها می‌توانند خوش‌خیم، بدخیم و دارای رفتاری نامشخص یا ناشناخته باشند [۲]. تومورهای خوش‌خیم

<sup>1</sup>Tumor

<sup>2</sup>Malignant tumor

<sup>3</sup>Cancer

<sup>4</sup>Tumor heterogeneity

<sup>5</sup>Oncogene

<sup>6</sup>Colorectal carcinoma

<sup>7</sup>Glioblastomas

شامل فیبروییدهای رحمی<sup>۸</sup> و خالهای ملانوسیتیک<sup>۹</sup> است. آنها محدود و محلی<sup>۱۰</sup> هستند و به سرطان تبدیل نمی‌شوند [۴]. تومورهای بالقوه بدخیم<sup>۱۱</sup> شامل سرطان در محل<sup>۱۲</sup> هستند. آنها به سایر بافت‌ها حمله نکرده و از بین نمی‌روند اما ممکن است به سرطان تبدیل شوند [۳]. تومورهای بدخیم را معمولاً سرطان می‌نامند. آنها به بافت اطراف حمله کرده و از بین می‌روند، ممکن است متاستاز<sup>۱۳</sup> 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<sup>۱۴</sup> و رگ‌زایی<sup>۱۵</sup> هستند. ناهمگنی تومور می‌تواند بین تومورها (ناهمگنی بین توموری) و یا درون تومورها (ناهمگنی درون توموری) رخ دهد. به طور گسترده‌ای پذیرفته شده است که توسعه تومور یک روند تکاملی است [۱۰]، و پیشرونده<sup>۱۶</sup> 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<sup>۱۷</sup> بوده که قابلیت ارث‌بری دارد. [۱۰]. 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 [۱۰]. بنابراین، برای توصیف و از بین بردن سلول‌های بدخیم در تومورها، لازم است که بر بخش کوچکی از سلول‌های تومورزا تمرکز کنیم [۳۰]. 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 در نهایت تومور حاصل میلیارد‌ها سلول بدخیم است که حاصل از تجمع تعداد زیادی جهش است [۲۷]. تکامل تومور به عنوان توالی پیدرپی گسترش کلونی توصیف می‌شود، که در آن در هر حالت جدید یک رویداد جهش اضافی ایجاد می‌شود [۱۰].

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

<sup>8</sup>Uterine fibroid

<sup>9</sup>Melanocytic nevi

<sup>10</sup>Local

<sup>11</sup>Potentially malignant tumor

<sup>12</sup>Carcinoma In Situ

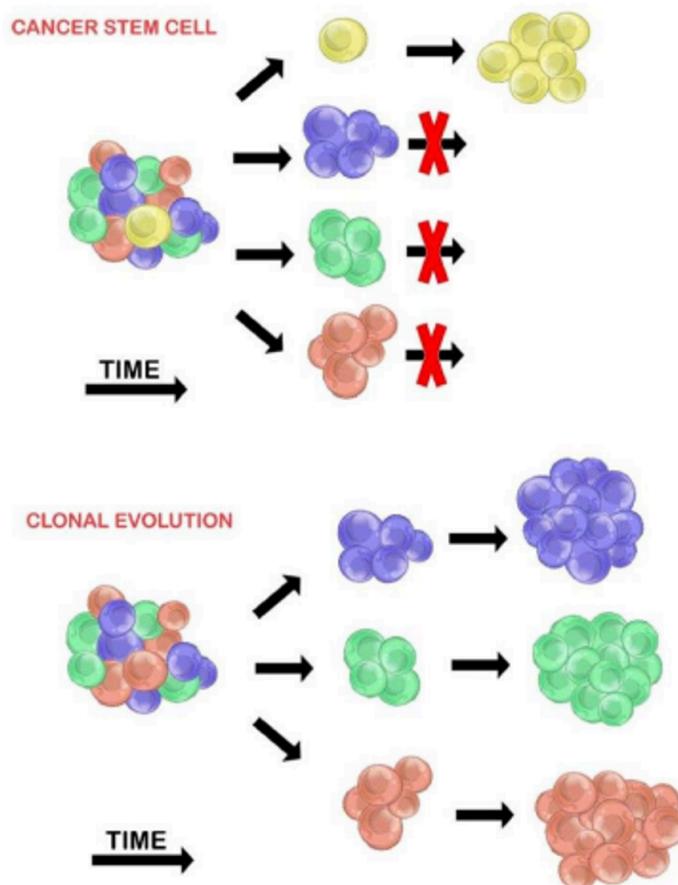
<sup>13</sup>Metastases

<sup>14</sup>Proliferative

<sup>15</sup>Angiogenic

<sup>16</sup>Spontaneous

<sup>17</sup>Clonal



شکل ۱.۱: دو مدل برای ناهمگونی تومور

[۱۰]. 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 [۳۴]. این مدل بیش از مدل خطی با ناهمگونی تومور مرتبط است. جهش‌های اکتسابی منجر به افزایش بی ثباتی ژنومی با هر نسل متوالی می‌شود [۱۴].

تومورهای ناهمگن<sup>۱۸</sup> که متشکل از چندین کلون هستند، می‌توانند حساسیت‌های مختلفی را نسبت به داروهای سمیت سلولی<sup>۱۹</sup> در نشان دهند. علاوه بر این، می‌زان ناهمگونی تومور می‌تواند خود به عنوان نشانگر زیستی<sup>۲۰</sup> مورد استفاده قرار گیرد زیرا هر چقدر می‌زان ناهمگونی تومور بیشتر باشد، احتمال حضور کلون‌های مقاوم در برابر درمان بیشتر است [۴۶]. دلایل حساسیت‌های مختلف می‌تواند تعاملات بین کلون‌ها باشد که ممکن است اثر درمانی را مهار یا تغییر دهد [۱۰]. تومورهایی با ناهمگونی زیاد، با احتمال بیشتری از کلون‌های گوناگون تشکیل

<sup>18</sup>Heterogenetic<sup>19</sup>Cytotoxic<sup>20</sup>Biomarker

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 [۲۰]. ناهمگنی‌های توموری یکی از عوامل اصلی مقاومت در برابر دارو است و بنابراین، یک عامل بالقوه در شکست درمان محسوب می‌شود. [۲۰]. 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 [۱۲]. بنابراین، ناهمگنی تومور می‌تواند برای درک توسعه تومور، پیچیدگی ایجاد کند و توسعه روش‌های موفقیت آمیز را با چالش روپرور کند [۲۰]. 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 دهنده [۱۲]. 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 توالی‌یابی، روش‌های مختلف ساخت درخت فیلوزنی 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 توالی‌یابی تکسولی، درخت فیلوزنی تومور را استباط کرده‌اند پرداخته شد. هر یک از این روش‌ها برای ساخت درخت فیلوزنی به همراه دادگان مورد استفاده، مورد ارزیابی قرار گرفت و در انتهای فصل سوم مقایسه‌های بین روش‌های مختلف صورت گرفت. در فصل چهارم روش پیشنهادی استباط درخت فیلوزنی بر مبنای یادگیری تقویتی و داده‌های توالی‌یابی 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 استتباط درخت تکاملی تومور معرفی می‌شوند. در پایان مفاهیم مرتبط با یادگیری ماشینی، یادگیری عمیق و یادگیری تقویتی به منظور استتباط درخت تکاملی تومور با رویکرد مبتنی بر داده<sup>۱</sup> توضیح داده می‌شوند.

### ۱.۲ تنوع ژنتیکی

دی‌ان‌ای<sup>۲</sup> یک مولکول بیولوژیکی است که توسط نوکلئوتیدها<sup>۳</sup> پلیمری شده است. در دی‌ان‌ای چهار نوع نوکلئوتید وجود دارد: آدنین<sup>۴</sup>، (A) تیمین<sup>۵</sup>، (T) سیتوزین<sup>۶</sup> (C) و گوانین<sup>۷</sup> (G). دی‌ان‌ای اساس توالی اسیدهای آمینه است که پروتئین را تشکیل می‌دهد. یک مولکول دی‌ان‌ای از دورشته تشکیل شده است. که در موازات<sup>۸</sup> هم و درجهت‌های مخالف قرار دارد و ساختاری از مارپیچ دوتایی ایجاد می‌کند. هر نوع نوکلئوتید روی یک رشته با نوع دیگری از نوکلئوتید در رشته دیگر مرتبط است: A با T؛ C با G (شکل ۱.۲) [۹]. این به عنوان قانون پایه

<sup>1</sup>Data driven

<sup>2</sup>DNA

<sup>3</sup>Nucleotid

<sup>4</sup>Adenine

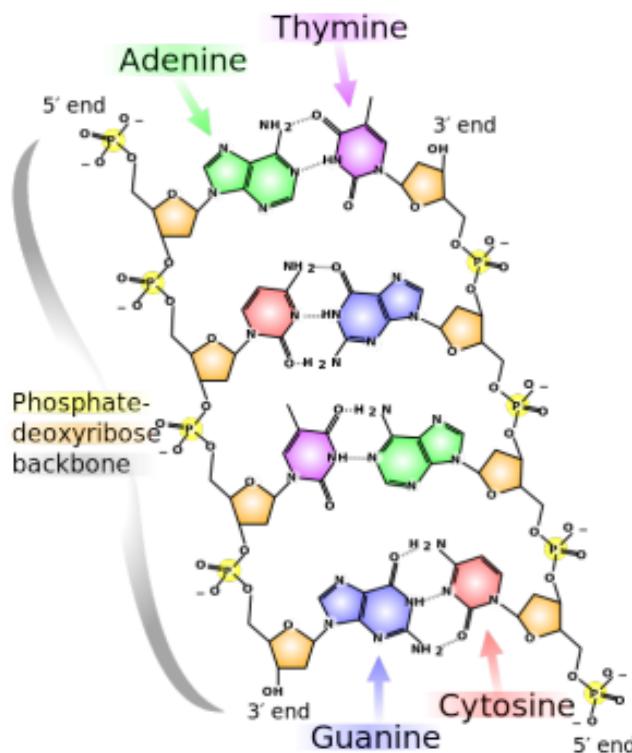
<sup>5</sup>Thymine

<sup>6</sup>Cytosine

<sup>7</sup>Guanine

<sup>8</sup>Antiparallel

جفت شدن نوکلئوتیدها در هر رشته از دی ان ای شناخته می شود.



شکل ۱.۲: مارپیچ دوگانه دی ان ای

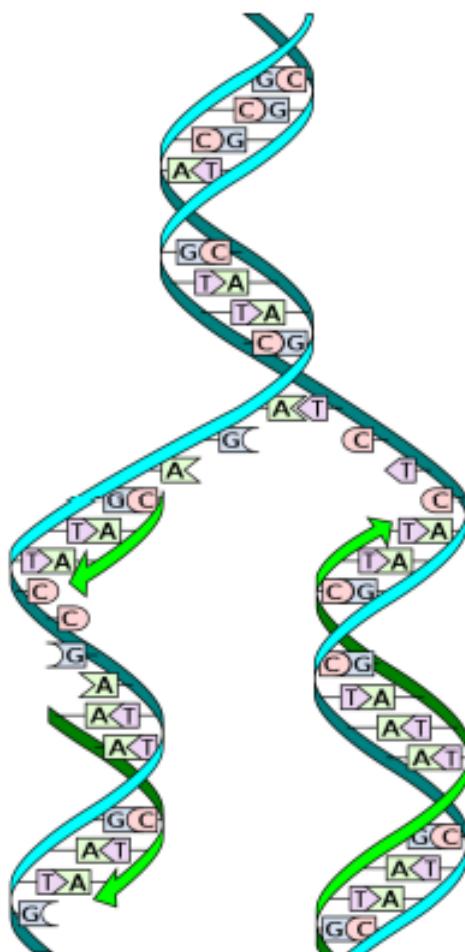
همانند سازی دی ان ای فرآیند تولید دو مولکول دی ان ای یکسان از مولکول دی ان ای اصلی است. وقتی تکثیر شروع می شود، دو رشته یک مولکول دی ان ای از یکدیگر جدا می شوند و هر رشته به عنوان الگویی برای ساخت نمونه مشابه خود عمل می کند. نوکلئوتیدها در هر موقعیت از یک رشته با نوع دیگری از نوکلئوتید مبتنی بر قانون 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 ناحیه‌ای از دی ان ای است و به عنوان مولکول واحد وراثت شناخته می شود. ژن‌های متعددی در ساختار دی ان ای با عملکردهای متفاوت وجود دارد. جهش به تغییر دائمی توالی هسته‌ای ژنوم اتصال می شود. جهش‌ها می توانند در حین فرآیند تکثیر دی ان ای و با جفت‌گیری اشتباه در قسمت‌های مختلف دی ان ای ایجاد می شود. انواع مختلفی از جهش‌ها مانند جهش تک نوکلئوتیدی<sup>۹</sup> (جهش نقطه‌ای)<sup>۱۰</sup> (شکل ۳.۲) و تغییرات ساختاری<sup>۱۱</sup>

<sup>9</sup>Single nucleotide mutation

<sup>10</sup>Point mutation

<sup>11</sup>Single variant



شکل ۲.۲: همانندسازی دی ان ای

شامل درج<sup>۱۲</sup>، حذف<sup>۱۳</sup> و برگشت<sup>۱۴</sup> (شکل ۴.۲) وجود دارد. جهش‌های سلولی می‌توانند به بنا بر دلایلی چون مواد شیمیایی، سمیت یا ویروس ایجاد شوند. جهش در یک ژن می‌تواند محصولات آن را تغییر دهد (مانند ایجاد پروتئین متفاوت) یا از عملکرد صحیح ژن جلوگیری کند [۶].

<sup>12</sup>Insertion<sup>13</sup>Deletion<sup>14</sup>reversion

original sequence:

ACTTGGTCAGAATTCCCAGGTGTCA

point mutation:

ACTTGGTCATAATTCCCAGGTGTCA

شکل ۳.۲: جهش تکنوکلئوتیدی

insertion:

ACTTGGTCAGAATTCCCAGGTGTCA  
↓  
ACTTGGTCAGATAGGCATTCCCAGGTGTCA

deletion:

ACTTGGTCAG~~GAATT~~CCCAGGTGTCA  
ACTTGGTCACCCAGGTGTCA

reversion:

ACTTGGTCAGAATTCCCAGGTGTCA  
X  
ACTTGGTC~~TTAAGA~~CCCAGGTGTCA

شکل ۴.۲: تغییرات ساختاری

## ۲۰.۲ تکامل تومور

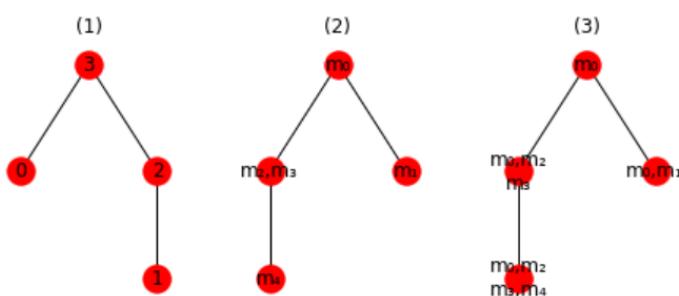
جهشی که در هر سلول از بدن اتفاق می‌افتد، به استثنای سلول‌های جنسی (اسپرم و تخمک)، جهش جسمی<sup>۱۵</sup> نامیده می‌شود [۱]. تجمع جهش بدنی در طول زندگی یک فرد می‌تواند منجر به رشد کنترل نشده مجموعه‌ای از سلول (تومور) شود [۳۶] و می‌تواند باعث شکل‌گیری سرطان یا بیماری‌های دیگر شود [۱]. بدلیل تجمع سلول‌های گوناگون، بیش از یک نوع سلول در تومور وجود خواهد داشت. به گروه‌های سلول با مجموعه‌ای از 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 [۱]. درخت فیلوزنیک رابطه تکاملی بین

<sup>15</sup>somatic

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 برچسب 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  $m_1, m_0$  دارد. کلون ۱ دارای جهش  $m_0, m_2$ ، کلون ۲ دارای جهش  $m_0, m_3$  و کلون ۳ دارای جهش  $m_0, m_1, m_2, m_3$  است.



شکل ۵.۲: درخت فیلوزنیک تومور

## ۳.۲ تکنولوژی‌های توالی‌یابی و فراوانی تغییرات آلل

تعیین توالی دی‌ان‌ای روشنی برای تشخیص ترتیب دقیق نوکلئوتیدها در یک رشته دی‌ان‌ای است. روش توالی‌یابی نسل بعدی<sup>۱۶</sup> 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<sup>۱۷</sup> گفته می‌شود). سپس خوانش با استفاده از الگوریتم هم‌تازی<sup>۱۸</sup> متنوعی مانند الگوریتم تبدیل Burrows-Wheeler با ژنوم مرجع تراز می‌شوند. پس از ترازبندی، می‌توان با جمع‌آوری خوانش‌های همپوشانی<sup>۱۹</sup>، توالی اجماعی<sup>۲۰</sup> 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

<sup>16</sup>Next generation sequencing

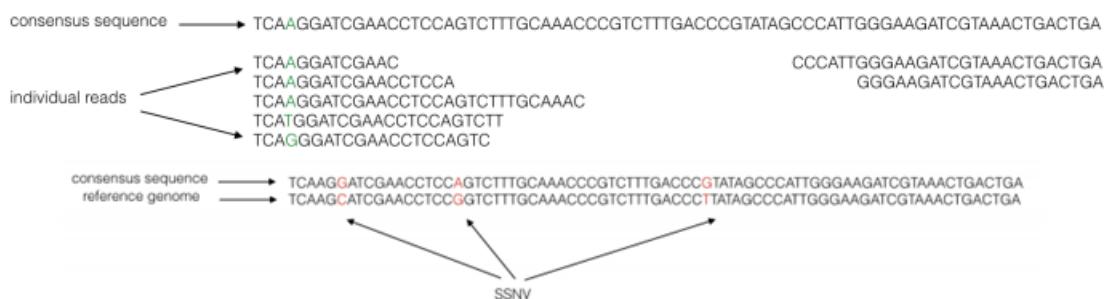
<sup>17</sup>Read

<sup>18</sup>Alignment

<sup>19</sup>Overlapping read

<sup>20</sup>Consensus

کل قرائت مرتبط با یک نوع جهش، را پوشش خوانش<sup>۲۱</sup> 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 مرتع هستند، شناسایی شده و به عنوان تغییرات بدنی تک نوکلئوتیدی<sup>۲۲</sup> 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 فیلوجنتیک 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

## ۴.۲ ناهمگنی ژنومی تومور

سرطان بیماری‌ای است که بدلیل ایجاد ناهنجاری‌های اساسی در فرآیندهای بنیادی سلول مانند تکثیر<sup>۲۳</sup>، تمایز<sup>۲۴</sup> و مرگ<sup>۲۵</sup> سلول ایجاد می‌شود [۲۶]. این ناهنجاری منجر به رشد کنترل نشده تومور و به کارگیری 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

<sup>21</sup>Read coverage

<sup>22</sup>Somatic single nucleotide variation

<sup>23</sup>Replication

<sup>24</sup>Differentiation

<sup>25</sup>Death

و منحصر به فرد است. این ژنوم که به ژنوم جوانهزنی<sup>۲۶</sup> 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 [۴۱]، 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 [۴۲]، قرار گرفتن در معرض جهش‌های با منشأ داخلی یا خارجی یا واردشدن توالی‌های دی‌ان‌ای با منشأ بیرونی بدلیل قرار گرفتن در معرض ویروس است [۴۳]. غالباً در سرطان، جهش‌های بدنی باعث ایجاد اختلال در روند تکثیر دی‌ان‌ای یا ترمیم آن می‌شوند و حتی جهش‌های بدنی بیشتری ایجاد می‌کنند [۴۴]. نظریه کلونی بودن سرطان [۳۶] 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<sup>۲۷</sup> 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 ندارند، به عنوان جهش‌های مسافر<sup>۲۸</sup> شناخته می‌شوند. شایان ذکر است که تعریف جهش راننده 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 راننده شود و عامل اصلی مقاومت در برابر درمان باشد. همچنین جهش‌ها را می‌توان بر اساس نوع تغییری که در دی‌ان‌ای ایجاد می‌شود، به طبقات متمایز تقسیم کرد. حذف و تغییر تک نوکلئوتیدها<sup>۲۹</sup> جهش‌هایی هستند که یک پایه در ژنوم را به پایه دیگری تغییر می‌دهند. ایندل<sup>۳۰</sup> 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 بدنی<sup>۳۱</sup> 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 اتلاق می‌شود که منجر به حذف یا کپی شدن قسمتی از ژنوم می‌شود. تغییرات شماره کپی<sup>۳۲</sup> نوعی تغییر ساختاری هستند که شامل وارونگی (وقتی

<sup>26</sup>Germline genome

<sup>27</sup>Driver mutation

<sup>28</sup>Passenger mutation

<sup>29</sup>Single nucleotide variants (SNV)

<sup>30</sup>Indel

<sup>31</sup>Single Somatic Mutation

<sup>32</sup>Copy number alteration

قسمت بزرگی از ژنوم معکوس شده باشد) و انتقال متعادل (جایی که دو بخش ژنومی مکان‌های خود را با یکدیگر تعویض می‌کنند) می‌باشند<sup>[۴۲]</sup>. این گونه‌های مختلف جهش مستقل از یکدیگر نیستند و می‌توانند در رابطه با یکدیگر اتفاق یافته باشند (به عنوان مثال یک جهش می‌تواند منجر به تقویت یک وارونگی شود).

تکنیک توالی‌بایی نسل بعدی این امکان را فراهم کرده است تا با صرف هزینه بسیار کم و با استفاده از یک نمونه توموری، توالی‌بایی از دی‌ان‌ای صورت پذیرد و همین امر منجر به تحول گستردگی در زمینه مطالعه تکامل تومور شده زیر امکان نمونه-برداری در تعداد بسیار بالا را 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.

تقریباً 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<sup>[۳۳]</sup> یا سلول‌های ایمنی نفوذی<sup>[۳۴]</sup> 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<sup>[۲۵]</sup> توالی‌بایی ژنوم و تغییرات شماره کپی را از طریق نمونه‌های مکانی مجزا استخراج شده از سرطان 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 نشده‌اند 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 فیلورژنی به آن‌ها نسبت داده شد. بازسازی درخت فیلورژنیک تومور این امکان را فراهم آورده تا سیر تکاملی تومور با استفاده از شاخه‌های مختلف درخت فیلورژنی 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"<sup>[۳۵]</sup>، حضور ITH را نیز نشان داد. در این مطالعه آنها توالی‌بایی کامل ژنوم را در عمق متوسط ۱۸۸X بر روی تومور پستان 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 دهنده که برخی از این خوش‌های مجزارانمی‌توان با جهش‌های موجود در تمام جمعیت‌های سرطانی توضیح داد، که این نشان دهنده

<sup>33</sup>Surrounding stroma

<sup>34</sup>Infiltrating immune cell

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. نویسندهای همچنین با تأیید اینکه جهش‌های زیر کلونال محدود به تغییرات تک نوکلئوتیدی نیستند، توانستند حضور تغییرات شماره کپی‌های کلونال و زیرکلونال را تأیید کنند. نویسندهای یک الگوریتم خوشه‌بندی غیرپارامتریک (یک مدل مخلوط فرآیند دیریشله (DPMM)) را با استدلال قابل توجه دستی برای استنباط 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 [۳]. 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 خطأ 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

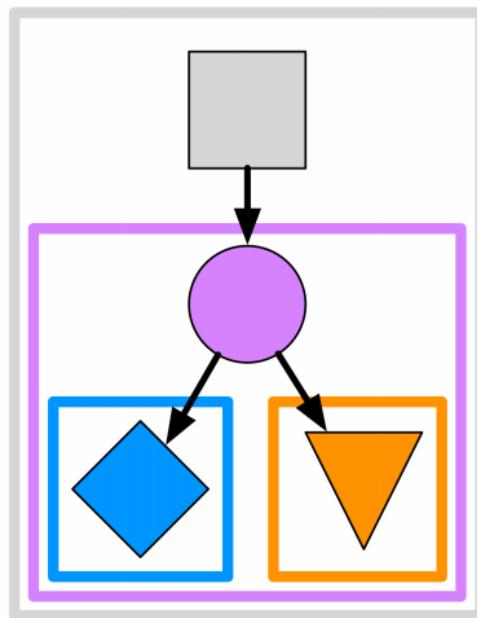
بازسازی ساب کلونال سعی دارد ژنوتیپ‌های موجود در تومور را از تعداد کمی از نمونه‌های توالی دی‌ان‌ای از آن تومور استنباط کند. تعداد ژنوتیپ‌های موجود در تومور از قبل مشخص نیست. این ژنوتیپ‌های زیر کلونال به طور معمول با جهش‌هایی که در مقایسه با ژنوم خط جوانه‌ای دارند، توصیف می‌شوند. ژنوم جوانه‌زنی علاوه بر نمونه‌های تومور، با تعیین توالی یک نمونه غیرسرطانی تعیین می‌شود. در حال حاضر در هنگام تعریف این جمعیت از دو نوع جهش به طور معمول استفاده می‌شود: جهش‌های ساده بدنی‌های متشکل از تعویض‌ها و درج / حذف کوچک (ایندل) و CNA 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 ایده آل، هر سلول در هر بخش یک جهش پیدا می کند [۹]، به همین ترتیب، بیشتر سلول‌های تومور ژنتیک منحصر به فردی خواهند داشت. بنابراین، به طور دقیق، اکثر سلول‌های تومور می‌توانند به طور بالقوه نمایانگر زیرجمعیت 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. بعلاوه، سلول‌هایی که بخش عمده‌ای از جهش‌های خود را به اشتراک می‌گذارند، خصوصاً جهش‌های راننده، صفات مشابهی دارند. به همین ترتیب، من قرارداد گسترهای را اتخاذ کرده و یک زیر جمعیت را به عنوان تمام سلول‌هایی که دارای زیر مجموعه یکسان جهش‌های بدنه در هنگام فراخوانی نوع هستند، تعریف می‌کنم.

یک گام مهم در بازسازی ساب کلونال محاسبه شیوع سلولی تبارهای زیرکلونال و سپس، در نهایت، زیرجمعیت‌های سرطانی است. شیوع سلولی یک زیرجمعیت، نسبت سلول‌های نمونه توالی شده متعلق به آن است. غالباً، شیوع سلولی با تقسیم بر خلوص نمونه، یعنی نسبت سلول‌های سرطانی در نمونه، به بخش سلول‌های سرطانی، نسبت سلول‌های سرطانی، تبدیل می‌شود. هر سلول دقیقاً یک زیرمجموعه تعلق دارد، بنابراین این شیوع باید در یک جمع باشد. به طور کلی، سلول‌های غیر سرطانی در یک زیرمجموعه واحد قرار می‌گیرند. با این حال، از آنجا که جهش‌ها اغلب در زیرجمعیت‌های متعدد وجود دارند، شیوع سلولی بسیاری از زیرجمعیت‌ها را نمی‌توان مستقیماً از جهش‌های آن استنباط کرد. برای پرداختن به این موضوع، ما یک نسب زیرکلونال برای یک جهش به عنوان مجموعه زیرجمعیت‌هایی که در آن وجود دارد، تعریف می‌کنیم. به طور رسمی، دودمان‌های زیرکلونال از زیر جمعیت بنیانگذار تشکیل می‌شود (جایی که جهش برای اولین بار ظاهر می‌شود) و همه زیرجمعیت‌های بعدی آن (که وراثت جهش) علاوه بر جهش‌های خاص خود، این زیرمجموعه‌های 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 سلولی یک تبار مجموع شیوع سلولی زیرجمعیت‌هایی است که متعلق به آن تبار هستند. از آنجا که سلول‌ها می‌توانند در چندین نژاد زیرکلونال 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. فلاش‌ها از جمعیت والدین به سمت فرزندانشان هدایت می‌شوند. دودمانهای زیرکلونال به صورت مستطیل نشان داده می‌شوند و با توجه به زیرمجموعه بنیادی آنها که در ریشه تیغه یافت می‌شوند، رنگی هستند.



شکل ۷.۲: درخت کلون تومور

## ۶.۲ تغییرات تعداد کپی

بیشتر ژنوم انسان دیپلوبتید است، به این معنی که دو نسخه از توالی دی ان ای ما در سلول های ما وجود دارد، یکی از پدر و دیگری از مادر. تغییرات شماره کپی این تغییر را می دهند، یا با تغییر در تعداد نسخه ها (مثلاً از طریق تکثیر کل ژنوم)، نسبت کپی های مادر به پدر (مثلاً از دست دادن خشی هتروزیگوتیه در تعداد کپی ها، جایی که برای همان منطقه یک ژنوم والدین تکثیر می شود و دیگری حذف شده است) یا هر دو (به عنوان مثال کپی کروموزوم مادر). بیشتر این تغییرات (به استثنای تکثیر کل ژنوم) دامنه محدودی از ژنوم را تحت تأثیر قرار می دهد، اما می تواند از تأثیر یک ژن تا یک کروموزوم کامل باشد. این بخش از ژنوم تغییر یافته به عنوان یک بخش شناخته می شود.

تغییرات شماره کپی می توانند تعداد کپی کل یک بخش و / یا تعداد 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 می توان 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 نمونه سرطانی ترسیم شده است. از نسبت نسبت ها برای محاسبه این واقعیت استفاده می شود که تعداد کل قرائت ها اغلب بین توالی یابی سرطانی و غیر سرطانی 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 دیپلولئید 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 تازمانی 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 [۱۳] رخ می‌دهد و بنابراین برای بخش‌های طولانی بسیاری از پولیمورفیسم تک هسته ای 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 [۲۵] 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  $\Phi$  فرض شده با ترکیب مناسب در تمام بخشها تعیین می‌شود. الگوریتم با بهینه سازی این تناسب بهترین مقدار  $\Phi$  را انتخاب می‌کند. سپس برای هر بخش، شماره کپی اصلی و جزئی با بهینه سازی متناسب بودن قطعه با بهترین مقدار  $\Phi$  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 کپیتخت کلونال هستند با جستجوی بخش‌هایی با اطلاعات مناسب

ضعیف با استفاده از  $\Phi_c$  استنباط شده مشخص می‌شوند. در این بخش‌ها، روش به طور همزمان و مستقل از هر بخش دیگر، عدد  $\Phi_i$  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 متداول 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 عمدۀ 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 [۴۷، ۴۹] و TITAN [۴۸]). تعداد دودمانهای زیر کلونال معمولاً با استفاده از احتمال جرمیمه 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 [۴۹]). 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 توالی‌یابی 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 ترازها 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 [۲۲]. 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) [۳۷]. 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 خطاهای توالی دقیقاً در همان موقعیت‌ها است. به نظر می‌رسد این خطاهای شیوع کم جهش‌های ساده

بدنی دارند. تمایز بین این خطاهای وشیوع کم واقعی جهش‌های ساده بدنه شامل یک معامله بین حساسیت و ویژگی و در حالت ایده‌آل، یک مدل نویز بسیار دقیق است. حل این مشکل امتداد طبیعی کار گسترهای است که در زمینه فراخوانی جهش‌های جوانهزنی انجام شده است و الگوریتم‌های زیادی برای انجام این کار وجود دارد (به عنوان مثال [۱۶، ۲۲])

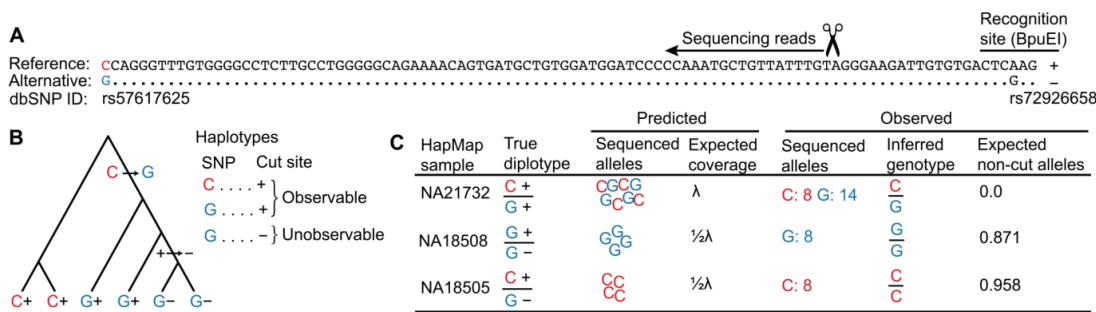
## ۸.۲ ترک آللی

اگرچه روش‌های تعیین توالی با بازدهی بالا [۳۲] ارزان هستند، اما تحت تاثیر مقدار بایاس هستند و مارکرهای ژنتیکی ای تولید می‌کنند که تقریباً به طور تصادفی در کل ژنوم تقسیم می‌شوند. این روشها با موفقیت در نگاشت صفات [۳۶، ۲۶]، ساخت مپ پیوندی [۱۹، ۳۸]، اسکن انتخاب [۱۷، ۴۸]، و برآورد تنوع ژنتیکی [۱۵] استفاده شده است. یکی از این روش‌ها، تعیین ژنوتیپ براساس توالی [۵] (GBS) است. در GBS، هدف توالی‌یابی فقط با اتصال آداتورهای توالی به محل‌های برش آنژیم 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 [۲۸]. 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 استفاده قرار می‌دهد که از توزیع با میانگین  $\lambda$  (آللیک‌های بدون برش صفر)، توزیع با میانگین  $\lambda/2$  (یک آللیک بدون برش)، یا با میانگین صفر (دو آللیک بدون برش). GBStools حداقل احتمال پارامتر  $\lambda$  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 حداقل‌رساندن 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

<sup>35</sup>Mapping

فرکانس سایت آلل‌های SNP مرجع قابل مشاهده و جایگزین،  $\varphi_1$  و  $\varphi_2$ ، و آلیک بدون برش،  $\varphi_3$ ، که در آن  $\varphi_1 + \varphi_2 + \varphi_3 = 1$  برآورد می‌کند و در نهایت، آزمون نسبت احتمال با مقایسه فرضیه صفر  $= \varphi_3$  با فرضیه  $> \varphi_3$  جایگزین می‌کند. GBStools در اجرای فعلی خود نمی‌تواند ژنتیپ‌های واقعی پنهان شده توسط کاوش آلیک را استنباط کند، اما می‌توان با فیلتر کردن سایت‌هایی که نسبت احتمال آنها زیاد است خطای 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 مشتق 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 کاوشی محاسبه شده توسط GBStools (0.958) 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. توزیع احتمال می‌تواند تابع 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 [۲۸]. رویکرد بیزی در یادگیری ماشین شامل استنباط احتمالی مقادیر پارامترهای منوط به مشاهدات است [۲۹]. 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

برای انجام استنتاج بیزی<sup>۳۶</sup>، 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:

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

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

<sup>36</sup>Bayesian

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 ماکوف مونت کارلو، تاز زمانی که بتوان  $p(x) \propto p$  را محاسبه کرد، نیازی به محاسبه  $p(x)$  نیست. این زنجیره ماکوف 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 ماکوف مونت کارلو الگوریتم Metropolis-Hastings (MH) است [۳۱]. 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:

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

می‌توان نشان داد که الگوریتم MH تعادل دقیق را برآورده می‌کند و از این رو،  $p(x)$  توزیع تعادل است [۱۱]. در 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 [۲۴]. 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  $\sigma_n$  است که  $\sigma$  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:

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

حاصل جمع بی نهایت محاسبه ESS را می‌توان با استفاده از برآوردگر پریودوگرام کوتاه تطبیقی Sokal [۳۹] تخمین زد.

## ۱۰.۲ یادگیری ماشین و یادگیری تقویتی

آنالیز داده‌های بالینی یک حوزه مهم تحقیقاتی در انفورماتیک، علوم کامپیوتر و پزشکی است که توسط محققان شاغل در دانشگاه‌ها، صنعت و مراکز بالینی انجام می‌شود. یکی از بزرگترین چالش‌ها در تعزیز و تحلیل داده‌های پزشکی، استخراج و تعزیز و تحلیل داده‌ها از تصاویر است. در چند سال اخیر روش‌های یادگیری ماشین انقلابی بزرگ در بینایی کامپیوتر<sup>۳۷</sup> به وجود آورده است که راه حل‌های جدید و کارآمدی را در مورد خیلی از مسائل و مشکلات موجود در آنالیز تصاویر که مدت زمان طولانی است حل نشده باقی مانده‌اند معرفی می‌کنند. برای اینکه این انقلاب وارد حوزه آنالیز تصاویر پزشکی شود شیوه و روش‌های اختصاصی ای باید طراحی شوند تا خاص بودن تصاویر پزشکی را در نظر گیرند. سیستم‌های کامپیوتری هوشمند چندین دهه است که در دنیا جایگاه برجسته‌ای پیدا کرده‌اند. در حال حاضر، به خاطر تکنیک‌های جدید هوش مصنوعی<sup>۳۸</sup>، قابلیت پردازش کامپیوتری بالا و رشد گسترده تصویربرداری و ذخیره‌سازی دیجیتالی داده، کاربرد هوش مصنوعی در حال انتقال به حوزه‌های گوناگون می‌باشد. در حوزه پزشکی، سیستم‌های هوش مصنوعی به منظور آشکارسازی بیماری، پیش‌بینی و به عنوان استراتژی پشتیبان در تصمیم‌گیری بالینی در حال توسعه، کاوش و ارزیابی هستند. در زمینه سرطان سینه<sup>۳۹</sup> از هوش مصنوعی به منظور آشکارسازی زودهنگام و تفسیر ماموگرام‌ها<sup>۴۰</sup> به منظور بهبود غربالگری سرطان پستان و کاهش تشخیص مثبت کاذب<sup>۴۱</sup> استفاده می‌شود و این امکان فراهم شده است تا متخصصانی مانند رادیولوژیست‌ها<sup>۴۲</sup> بتوانند بر اساس میلیون‌ها تصویر از بیماران قبلی که مشخصات مشابهی دارند، تصمیمات آگاهانه‌ای بگیرند. استفاده از هوش مصنوعی در شیوه‌های تشخیص سرطان سینه به مدالیته تصویربرداری<sup>۴۳</sup> و همچنین تفسیر آسیب‌شناسی<sup>۴۴</sup> نیز گسترش یافته است. یادگیری عمیق<sup>۴۵</sup> که زیر شاخه‌ای از یادگیری ماشین می‌باشد یکی از تکنیک‌های هوش مصنوعی است که در انواع مختلفی از مسائل کلینیکی و پردازش تصاویر

<sup>37</sup>Computer Vision

<sup>38</sup>Artificial Intelligence (AI)

<sup>39</sup>Breast cancer

<sup>40</sup>Mammogram

<sup>41</sup>False positive

<sup>42</sup>Radiologist

<sup>43</sup>Imaging modality

<sup>44</sup>Pathology

<sup>45</sup>Deep learning

پزشکی شامل آشکارسازی<sup>۴۶</sup>/شناسایی<sup>۴۷</sup>، قطعه‌بندی<sup>۴۸</sup> و تشخیص به کمک کامپیوتر<sup>۴۹</sup> به کار گرفته می‌شود. یادگیری عمیق مجموعه‌ای از الگوریتم‌های ماشین است که قادر به مدل‌سازی الگوهای بطور مستقیم از داده‌های خام می‌باشد. الگوریتم‌های یادگیری عمیق از مجموعه‌ای از لایه‌های چندگانه با واحدهای پردازنده غیرخطی برای استخراج و تبدیل ویژگی استفاده می‌کنند. هر لایه از خروجی لایه قبل به عنوان ورودی استفاده می‌کند. این مفهوم با بسیاری از روش‌های دیگر یادگیری ماشین که نیاز به استخراج ویژگی دارند متفاوت است. به 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<sup>۵۰</sup>. 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<sup>۵۱</sup> تشکیل شده است که در یک سیستم پیچیده سازمان یافته‌اند. هر نورون بر اساس ورودی‌های خود، یک خروجی (فعال‌سازی<sup>۵۲</sup>) را محاسبه می‌کند که می‌تواند فعالیت‌ها یا داده‌های سایر نورون‌ها باشد. متدائل‌ترین نوع شبکه عصبی، شبکه عصبی کاملاً متصل شبکه عصبی کاملاً متصل پیش‌خور<sup>۵۳</sup> است. این شبکه‌ها دارای ورودی (جایی که داده‌ها وارد می‌شوند) و خروجی هستند. به طور معمول، هدف از استفاده از این مدل‌ها حل رگرسیون<sup>۵۴</sup> یا طبقه‌بندی<sup>۵۵</sup>، توسط تقریب فعال‌سازی خروجی با مقدار هدف، برای هر داده ورودی است. این شبکه‌ها به صورت لایه<sup>۵۶</sup> متواالی 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 غیرخطی عبور می‌دهد

محاسبه خروجی نورون  $i$  ام لایه  $k$

$$O_{k,i} = \text{actv}(\mathbf{W}_{k,i} \cdot \mathbf{l}_{k-1} + b_{k,i}) \quad (5.2)$$

<sup>46</sup>Detection

<sup>47</sup>Recognition

<sup>48</sup>Segmentation

<sup>49</sup>Computer-aided diagnosis

<sup>50</sup>Neuron

<sup>51</sup>Activation

<sup>52</sup>Fully-connected feed forward neural network

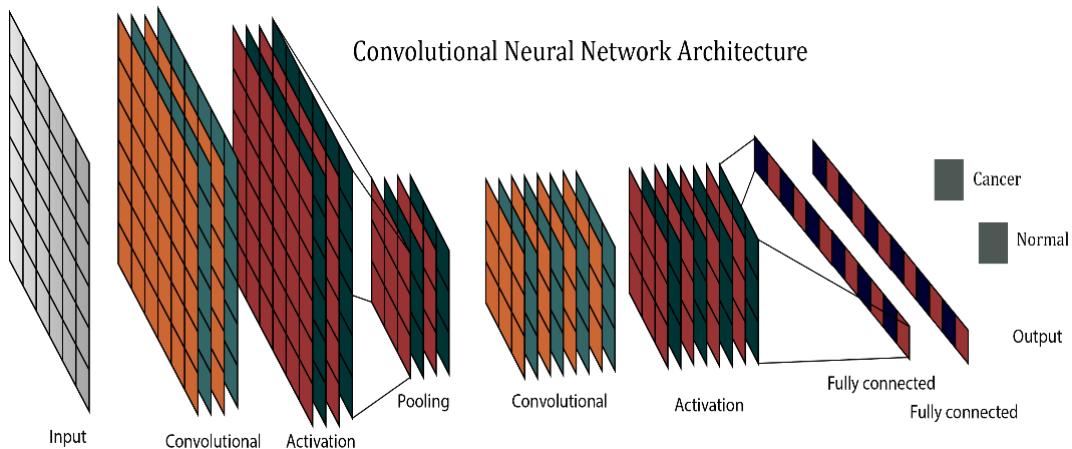
<sup>53</sup>Regression

<sup>54</sup>Classification

<sup>55</sup>Layer

که واحد  $O_{k,i}$  واحد  $i$  ام لایه  $k$  و  $b_{k,i}$  پارامترهای بردار تمام فعال‌سازهای لایه  $1 - k$  است. بردار  $W_{k,i}$  عدد  $l_{k-1}$  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 پارامترهای کافی داشته باشد می‌تواند هر تابع پیوسته ای را با خطأ دلخواه تقریب بزند [۱۸].

شبکه‌های عصبی کانولوشنی 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 [۴۴]. 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 سطح 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 [۲۳]. 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 گردیدی به تعویق افتاد [۳۳].



شکل ۹.۲: معماری یک شبکه عصبی کانولوشنی

همانطور که قبلاً 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

<sup>۵۶</sup>Network weight

تمام لایه‌های یک شبکه کانولوشنی و توابع مورد استفاده در آموزش آن‌ها را شرح می‌دهیم. یک معماری می‌تواند 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 رایج‌ترین حالت، یک تصویر به یک آرایه  $3 \times 3$ <sup>۵۷</sup> بعدی تبدیل می‌شود با ابعاد  $[w, h, 3]$  که  $w$  و  $h$  عرض و ارتفاع هستند. بعد آخر به دلیل استفاده از تصاویر رنگی <sup>۵۸</sup> RGB اغلب  $3 \times 3$  است. وقتی از تصاویر اشعه ایکس استفاده می‌کنیم چون دارای یک کanal <sup>۶۰</sup> 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 <sup>۶۱</sup> از این لایه به دست می‌آید. در نتیجه تمامی نوروون‌ها در یک نقشه ویژگی، وزن‌ها و بایاس‌ها <sup>۶۳</sup> 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 نوروون در نخستین لایه مخفی به ناحیه کوچکی از نوروون‌های ورودی متصل می‌شود. برای مثال اگر این ناحیه  $5 \times 5$  باشد این ناحیه کوچک  $25 \times 25$  پیکسلی ناحیه ادراک محلی <sup>۶۴</sup> یا کرنل کانولوشن نامیده می‌شود. با توجه به شکل <sup>۱۰.۲</sup> یک تصویر ورودی  $28 \times 28$  داریم که یک کرنل  $5 \times 5$  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 <sup>۱۰.۲</sup> مشخص است لایه مخفی شامل یک شبکه  $24 \times 24$  نورونی خواهد بود.

در شکل <sup>۱۰.۲</sup> هر نوروون لایه مخفی دارای یک بایاس و تعداد  $5 \times 5$  وزن می‌باشد که به ناحیه ادراکی خود متصل شده است. تمامی نوروون‌های لایه مخفی مذکور که دارای ابعاد  $24 \times 24$  هستند، دارای وزن‌ها و بایاس‌های مشترکی می‌باشند. به عبارت دیگر خروجی نوروون لایه کانولوشن  $y_{w,h,m}$  در طول و عرض  $w, h$  و عمق  $m$  به صورت رابطه <sup>۶۰.۲</sup> است.

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

<sup>57</sup> Dimension

<sup>58</sup> Red Green Blue

<sup>59</sup> X-ray

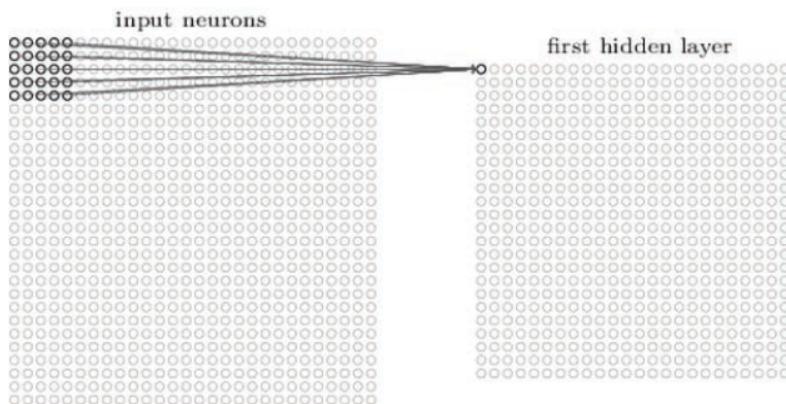
<sup>60</sup> Channel

<sup>61</sup> Intensity

<sup>62</sup> Feature map

<sup>63</sup> Bias

<sup>64</sup> Local receptive field



شکل ۱۰.۲: عملیات کانولوشن در یک شبکه عصبی کانولوشنی با کرنل  $5 \times 5$

که در این رابطه  $f$  تابع فعالیت،  $b_m$  بایاس مشترک نورون‌ها،  $W_{k,m}$  وزن‌های  $5 \times 5$  مشترک نورون‌ها و  $x_{i,j,k}$  ورودی در موقعیت  $k$ ،  $i, j$  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  $D_2 \times H_2 \times W_2$  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 نشان‌دهنده اندازه کرنل، مدار لایه‌گذاری صفر<sup>۶۵</sup>، اندازه اندازه گام<sup>۶۶</sup> و تعداد فیلترها می‌باشد. طبق این روابط به ازای هر فیلتر تعداد  $D_1 \times F \times F$  وزن داریم و با توجه به تعداد  $k$  فیلتر موجود، در مجموع تعداد  $(D_1 \times F \times F) \times k$  وزن و  $k$  بایاس ایجاد می‌شود. بنابراین تعداد پارامترهایی که شبکه در یک لایه کانولوشن خود می‌بایست آموخته بینند 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<sup>۶۷</sup> متداول ترین تابع مورد

<sup>65</sup>Zero padding

<sup>66</sup>Stride

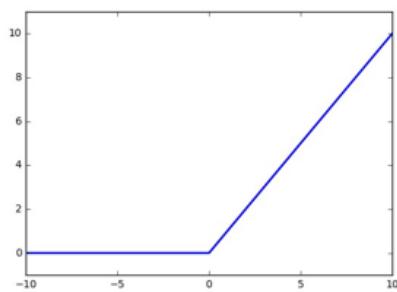
<sup>67</sup>Rectified linear unit (ReLU)

استفاده به خاطر آسان کردن مرحله آموزش است. مثال‌های دیگر شامل تابع سیگموید و هایپربولیک<sup>۶۸</sup> است.

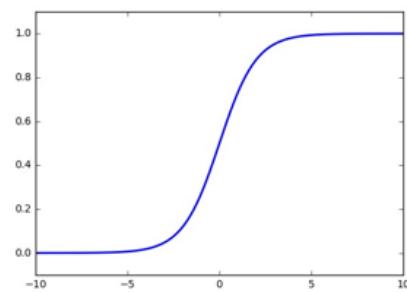
$$\text{ReLU: } r_{m,n,c} = \max\{\circ, l_{x,y,z}\} \quad (8.2)$$

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

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



(a) ReLU



(b) Sigmoid

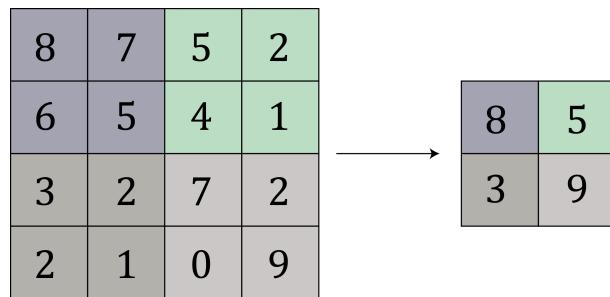
شکل ۱۱.۲: (a) تابع فعالیت ReLU و (b) تابع فعالیت سیگموید

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 بیش‌پردازش<sup>۶۹</sup> 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 mean-pooling و max-pooling نام دارند که تابع اول دارای کاربرد بیشتری در شبکه‌های عصبی کانولوشنی است. طریقه عملکرد max-pooling به این صورت است که در هر پنجره بزرگترین پیکسل<sup>۷۰</sup> 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.

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

<sup>68</sup>Hyperbolic tangent<sup>69</sup>Over-fitting<sup>70</sup>Pixel

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 ۲ 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 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

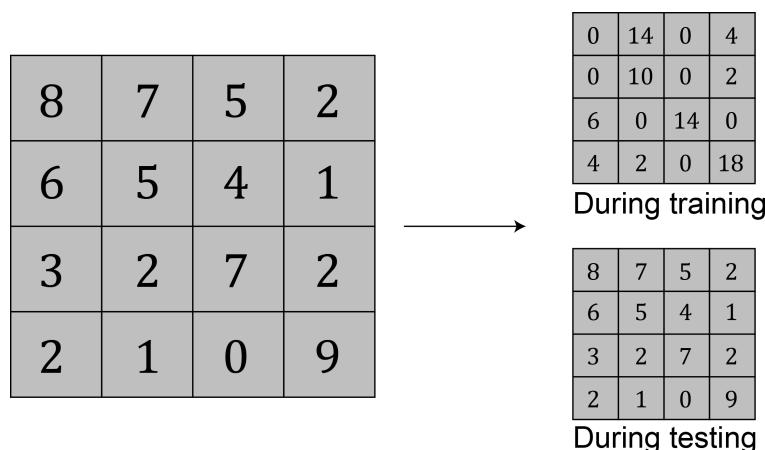
$$P(C)_j = \frac{e^{c_j}}{\sum_{k=1}^K e^{c_k}} \quad (10.2)$$

اگر دو کلاس داشته باشیم می‌توانیم از تابع SoftMax با دو خروجی استفاده کنیم یا از یک نورون استفاده کنیم و تابع سیگموید را محاسبه کنیم. برای دو کلاس احتمال توسط معادله؟؟ محاسبه می‌شود

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

حذف تصادفی یک روش بسیار رایج برای جلوگیری از بیش‌پردازش شبکه عصبی مصنوعی از جمله مدل‌های یادگیری عمیق است [۴۰]. 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 فعال سازی است. این مازول بطور تصادفی برخی از نقاط نقشه ویژگی ورودی را صفر می‌کند. هریک از مازول‌ها دارای یک احتمال مستقل  $\sigma$  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  $\sigma$  دارد، که برای

آموزش در فاصله [۱، ۰] 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 تصور کرد. مکانیزم حذف، نورون‌هایی را که به اتصالات ورودی کمتری متکی هستند را در نظر می‌گیرد. زیرا افت بک 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<sup>۱</sup> یک تکنیک جدید ولی خیلی کارآمد است. در طی آموزش مدل‌های عمیق، وزن‌ها در هر



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

تکرار<sup>۱۲</sup> 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<sup>۱۳</sup> می‌گویند. این پدیده فرایند آموزش را کند می‌کند، به مقدار دهی دقیق‌تر وزن احتیاج دارد و مانع بهینه‌سازی<sup>۱۴</sup> 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<sup>۱۵</sup> داخلی، این لایه برای هر دسته آموزش با کم کردن میانگین و تقسیم بر انحراف استاندارد<sup>۱۶</sup> 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 با یا بدون نرمال‌سازی دسته عادی نشان دهد، دو وزن جدید قابل تمرین<sup>۷</sup>

<sup>۱۱</sup>Batch normalization

<sup>۱۲</sup>Iteration

<sup>۱۳</sup>Internal covariate shift

<sup>۱۴</sup>Optimization

<sup>۱۵</sup>Covariance

<sup>۱۶</sup>Standard deviation

و  $\beta$  اضافه می‌شوند که خروجی را اندازه‌گیری و جبران می‌کنند. بنابراین خروجی به صورت معادله ۱۲.۲ است.

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

که  $u_c$  و  $v_c$  متوسط‌های در حال اجرا ( $I_c$ ) و  $\text{mean}(I_c)$  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 رگولاژیشن<sup>۷۷</sup> است. مدل با استفاده از تابع هزینه<sup>۷۸</sup> 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 خطأ در پیش‌بینی را کاهش دهد.

بهینه سازی 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  $\{0, 1\}$  هستند. 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<sup>۷۹</sup> SGD و Adadelta.

به حداقل رساندن تابع هزینه با کاهش گرادیان تقریباً رایج ترین الگوریتم برای بهینه سازی شبکه‌های عصبی است. اگر تابع هزینه آنتروپی متقاطع دودویی<sup>۸۰</sup> 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.

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

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

<sup>77</sup>Regularization

<sup>78</sup>Cost function

<sup>79</sup>Stochastic gradient descent

<sup>80</sup>Binary cross-entropy

بکپروپگیشن<sup>۱۱</sup> اساس آموزش شبکه عصبی است. این عمل تنظیم-دقیق وزن‌های یک شبکه عصبی بر اساس میزان خطا<sup>۱۲</sup> در هر دوره<sup>۱۳</sup> قبلی است که این امر با محاسبه مشتق‌های تابع خطا بر اساس وزن‌ها  $\nabla_{\theta} L(\theta)$  در زمان آموزش امکان پذیر است. تنظیم مناسب وزن‌ها باعث کاهش میزان خطا می‌شود. در فرایند بکپروپگیشن ابتدا ورودی در سراسر شبکه انتشار داده می‌شود سپس  $L(\theta)$  محاسبه شده و در نهایت این خطا از طریق تمام وزن‌ها در شبکه رو به عقب منتشر می‌شود. مشتق تابع هزینه از خروجی توسط معادله<sup>۱۴.۲</sup> محاسبه می‌شود.

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

همچنین محاسبه مشتق تابع هزینه  $L$  از ورودی  $i$  به صورت معادله<sup>۱۵.۲</sup> محاسبه می‌شود.

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

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 بهتر است مروی 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 ادرارک ماشینی، برچسب زدن یا خوشه بندی 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

<sup>۱۱</sup>Back-propagation

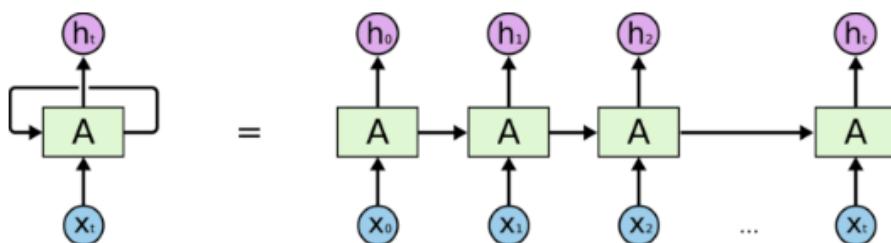
<sup>۱۲</sup>Loss

<sup>۱۳</sup>epoch

عصبي مصنوعی از تعداد زیادي مؤلفه‌ی پردازشی (نورون) تشکیل شده‌اند که اتصالات زیادي بینشان وجود دارد و برای حل یک مسئله با یکدیگر همکاری دارند. شبکه‌ی عصبي مصنوعی معمولاً تعداد زیادي 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 می‌کنند؛ 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.

## ۱۰.۱۱.۲ شبکه عصبي بازگشتی چیست؟

شبکه‌ی عصبي بازگشتی شکلی از شبکه‌ی عصبي پیشخور است که یک حافظه‌ی داخلی دارد. شبکه عصبي بازگشتی ذاتاً بازگشتی 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 بازگشتی فرستاده می‌شود. این شبکه برای تصمیم‌گیری، هم ورودی فعلی و هم خروجی که از ورودی قبلی آموخته شده را در نظر می‌گیرد. شبکه عصبي بازگشتی برخلاف شبکه‌های عصبي 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 عصبي، ورودی‌ها از یکدیگر مستقل هستند، اما در شبکه عصبي بازگشتی ورودی‌ها به هم مرتبط می‌باشند. به شکل ۱۴.۲ توجه کنید، این شبکه ابتدا  $X_1$  را از دنباله‌ی ورودی‌ها گرفته و خروجی  $h_1$  را تولید می‌کند که همراه با  $X_1$  ورودی گام بعدی



An unrolled recurrent neural network.

شکل ۱۴.۲: یک نمونه بازشده شبکه عصبي بازگشتی

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 عصبي بازگشتی بدین طریق می‌تواند هنگام آموزش زمینه را به خاطر داشته باشد.

فرمول حالت<sup>۱۴</sup> کنونی به صورت رابطه ۱۷.۲ خواهد بود که در آن،

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

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،  $W_{hh}$  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  $\text{ReLU}$  یا  $\tanh$

<sup>84</sup>State

## ۴.۱۱.۲ کاربردهای شبکه عصبی بازگشته

- شرح نویسی عکس<sup>۸۵</sup>: شبکه عصبی بازگشته با تحلیل حالت کنونی عکس، برای شرح نویسی عکس به کار می‌رود
- پیش‌بینی سری‌های زمانی<sup>۸۶</sup>: هر مسئله سری زمانی مانند پیش‌بینی قیمت یک سهام در یک ماه خاص، با شبکه عصبی بازگشته قابل انجام است
- پردازش زبان طبیعی<sup>۸۷</sup>: کاوش متن و تحلیل احساسات می‌تواند با استفاده از شبکه عصبی بازگشته انجام شود
- ترجمه ماشینی<sup>۸۸</sup>: 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.

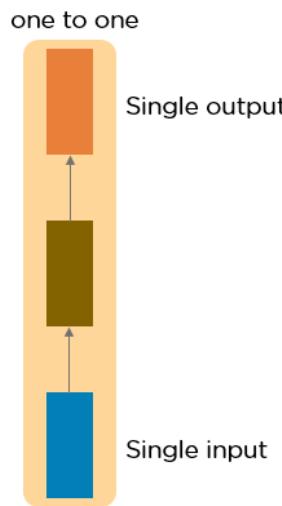
چند به چند (many to many): دنباله ایی از ورودی‌ها را می‌گیرد و دنباله ایی از خروجی‌ها را تولید می‌کند. ترجمه ماشینی نمونه ایی از این نوع شبکه است.

<sup>85</sup>Image Captioning

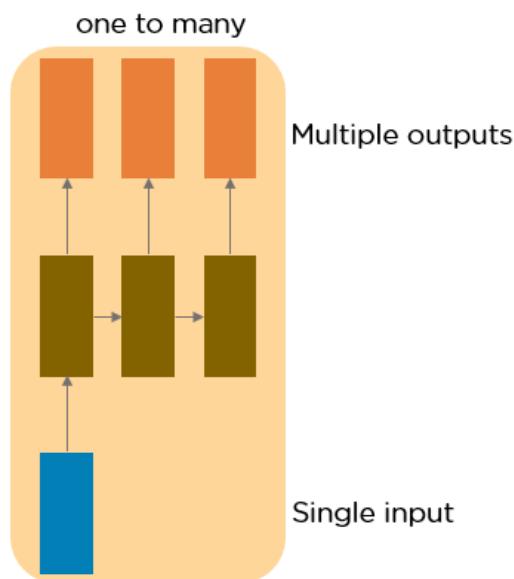
<sup>86</sup>Time Series Prediction

<sup>87</sup>Natural Language Processing

<sup>88</sup>Machine Translation



شکل ۱۵.۲: ساختار شبکه عصبی بازگشتی یک به یک

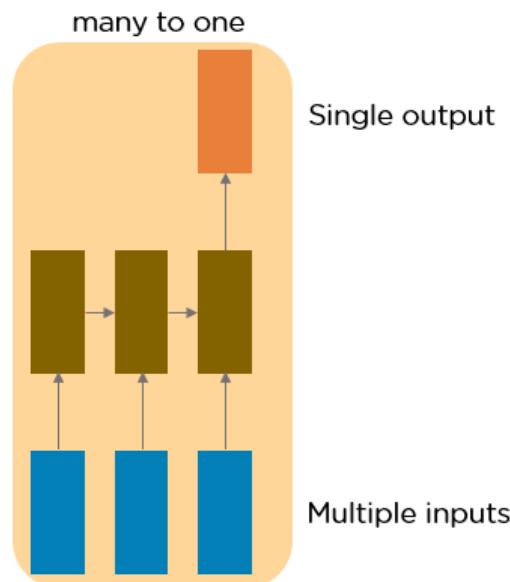


شکل ۱۶.۲: ساختار شبکه عصبی بازگشتی یک به چند

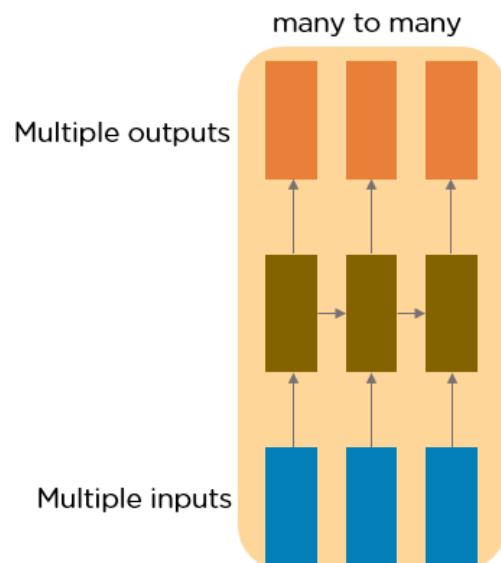
### ۶.۱۱.۲ حافظه‌ی کوتاه‌مدت بلند (LSTM)

شبکه‌های حافظه‌ی کوتاه‌مدت بلند<sup>۸۹</sup> 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 داشت نیز در این شبکه‌ها حل شده است. شبکه‌های LSTM برای مسائل رده‌بندی، پردازش و پیش‌بینی سری‌های

<sup>89</sup>Long Short Term Memory (LSTM)



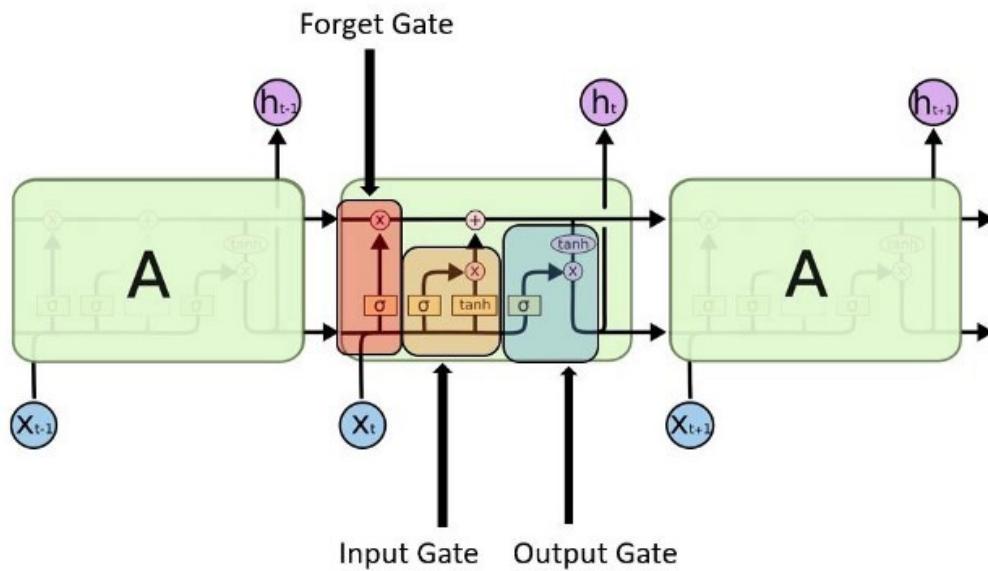
شکل ۱۷.۲: ساختار شبکه عصبی بازگشتی چند به یک



شکل ۱۸.۲: ساختار شبکه عصبی بازگشتی چند به چند

زمانی با استفاده از برقسپهای 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 ضریب دهی (بین  $-1$  تا  $+1$ ) 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 محتوا  $X_t$ ، 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.

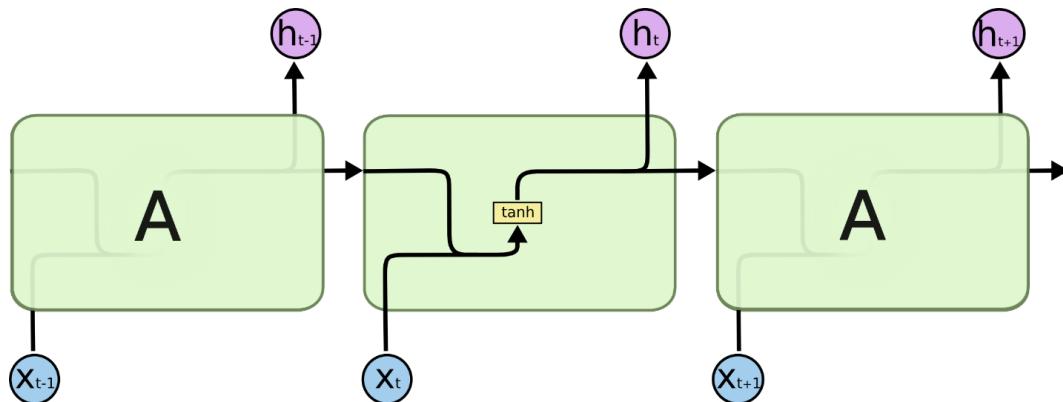
$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (21.2)$$

۳) دریچه‌ی خروجی: ورودی و حافظه‌ی بلوک برای تصمیم‌گیری در مورد خروجی مورد استفاده قرار می‌گیرند. تابع سیگموئید تصمیم می‌گیرد مقادیر بین ۰ و ۱ اجازه‌ی ورود دارند و تابع  $\tanh$  با ضریب‌دهی (بین -۱ تا +۱) به مقادیر و ضرب آن‌ها در خروجی تابع سیگموئید در مورد اهمیت آن‌ها تصمیم‌گیری می‌کند.

$$\begin{aligned} o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t &= o_t * \tanh(C_t) \end{aligned} \quad (22.2)$$

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.

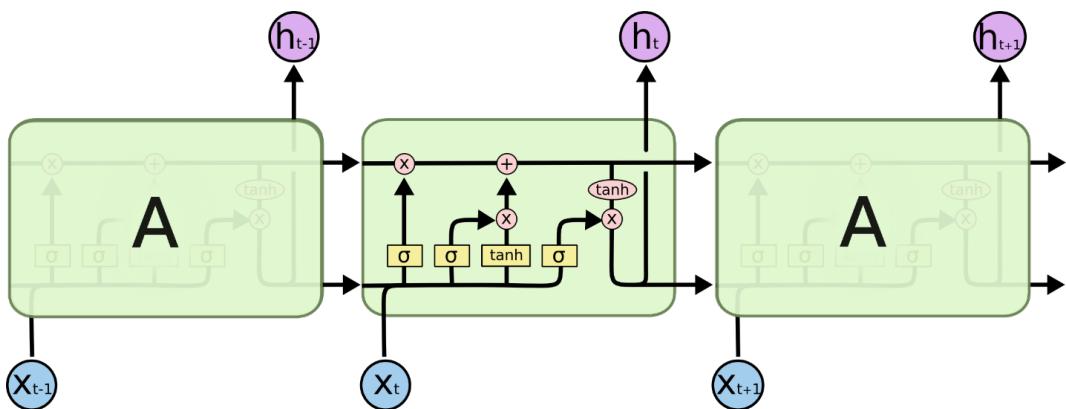
همه شبکه‌های عصبی بازگشتی به شکل دنباله‌ای (زنگیره‌ای) تکرار شونده از مازول‌های (واحدهای) شبکه‌های عصبی هستند. در شبکه‌های عصبی بازگشتی استاندارد، این مازول‌های تکرار شونده ساختار ساده‌ای دارند، برای مثال تنها شامل یک لایه تائزانتِ هایپربولیک ( $\tanh$ ) هستند.



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

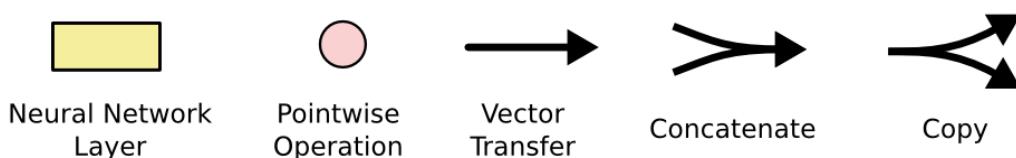
شبکه‌های LSTM نیز چنین ساختار دنباله یا زنجیره‌مانندی دارند ولی مازولِ تکرار شونده ساختار متفاوتی دارد. به جای داشتن تنها یک لایه شبکه عصبی، ۴ لایه دارند که طبق ساختار ویژه‌ای با یکدیگر در تعامل و ارتباط

هستند. در ادامه قدم به قدم ساختار شبکه‌های حافظه‌ی کوتاه‌مدت بلند را توضیح خواهیم داد. اما در ابتدا معنی



شکل ۲۱.۲: مراحلهای 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.

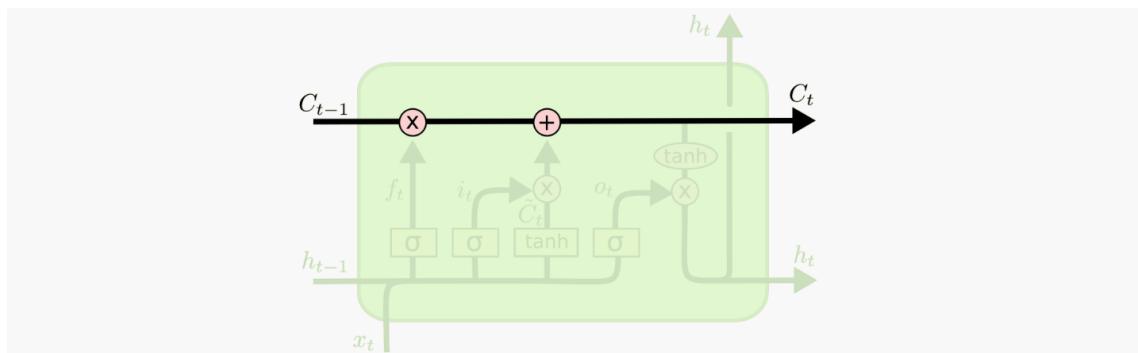
یک بردار را به صورت کامل از خروجی یک گره به ورودی گره دیگر انتقال می‌دهد. دایره‌های صورتی نمایش دهنده عملیات‌های نقطه به نقطه مانند «جمع کردن دو بردار» هستند. مستطیل‌های زرد، لایه‌های شبکه‌های عصبی هستند که شبکه پارامترهای آن‌ها را یاد می‌گیرد. خط‌هایی که با هم ادغام می‌شوند نشان‌دهنده الحاق<sup>۹۰</sup> 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<sup>۹۱</sup> است که در حقیقت یک خط افقی است که در بالای شکل ۲۲.۲ قرار 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).

LSTM این توانائی را دارد که اطلاعات جدیدی را به سلول حالت اضافه یا اطلاعات آن را حذف کنید. این کار

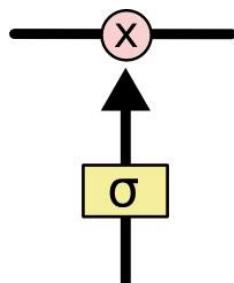
<sup>90</sup>Concatenation

<sup>91</sup>Cell state



شکل ۲۳.۲: سلول حالت در مازول LSTM

توسط ساختارهای دقیقی به نام دروازه‌ها<sup>۹۲</sup> 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

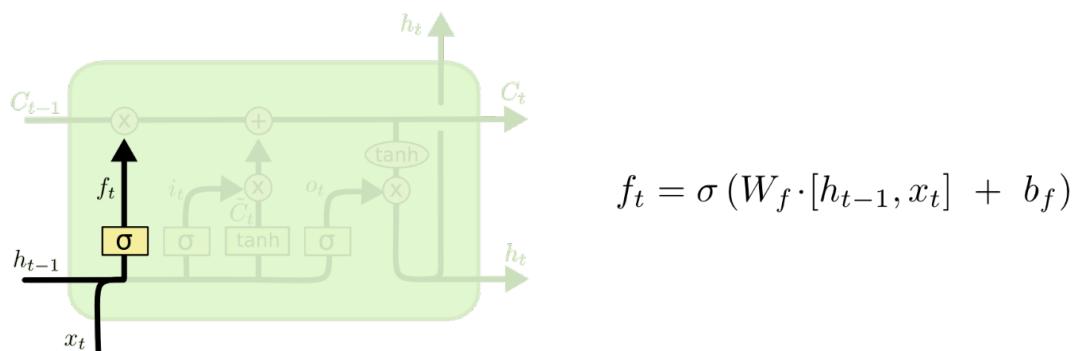
خروجی لایه سیگموید عددی بین صفر و یک است، که نشان می‌دهد چه مقدار از ورودی باید به خروجی ارسال شود. مقدار صفر یعنی هیچ اطلاعاتی نباید به خروجی ارسال شود در حالی که مقدار یک یعنی تمام ورودی به خروجی ارسال شود!

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.

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 «دوازه فراموشی»<sup>۹۳</sup> انجام می‌شود. این دروازه با توجه به مقادیر  $h_{t-1}$ ,  $x_t$ ,  $C_{t-1}$ , برای هر عدد، مقدار صفر یا یک را در سلول حالت  $C_t$  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

<sup>92</sup>Gate<sup>93</sup>Forget gate

کنونی  $C_{t-1}$  را پاک شود و هیچ مقداری از آن به  $C_t$  برد نشود. باید 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

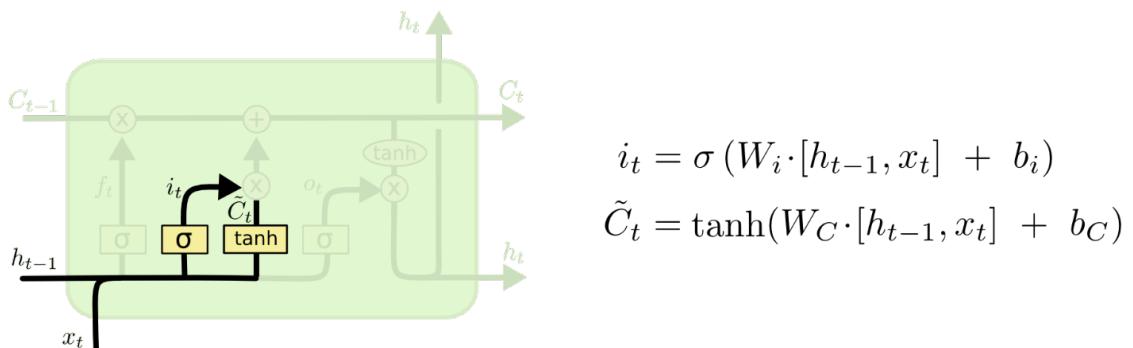
قدم بعدی این است که تصمیم بگیریم چه اطلاعات جدیدی را می‌خواهیم در سلول حالت ذخیره کنیم. این تصمیم دو بخشی است. ابتدا یک لایه سیگموید به نام دروازه ورودی<sup>۹۴</sup> داریم که تصمیم می‌گیرد چه مقادیری به روز خواهند شد. مرحله بعدی یک لایه تانژانت هایپربولیک است که برداری از مقادیر به نام  $\tilde{C}_t$  می‌سازد که می‌توان آن 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.

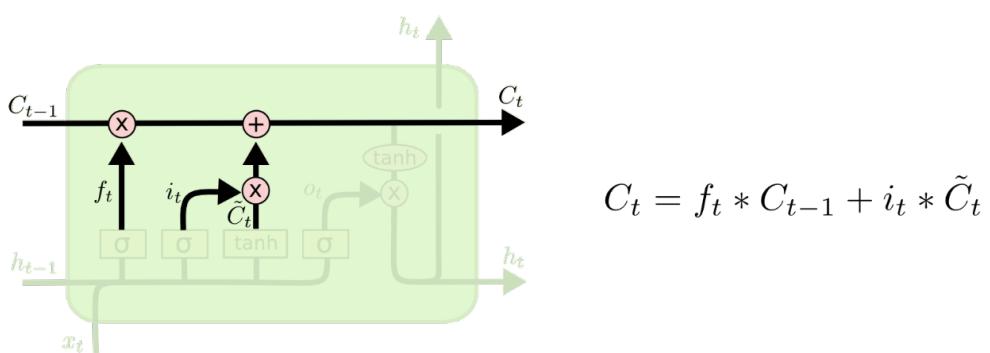
حال زمان آن فرا رسیده است که سلول حالت قدیمی یعنی  $C_{t-1}$  را سلول حالت جدید یعنی  $C_t$  به روز کنیم. در مراحل قبلی تصمیم گرفته شد که چه کنیم و در حال حاضر تنها لازم است تصمیماتی را که گرفته شد عملی کنیم. ما مقدار قبلی سلول حالت را در  $f_t$  ضرب می‌کنیم که یعنی فراموش کردن اطلاعاتی که پیش‌تر تصمیم گرفتیم آن‌ها را فراموش کنیم. سپس  $i_t * \tilde{C}_t$  را به آن اضافه می‌کنیم. در حال حاضر مقادیر جدید سلول حالت با توجه به تصمیماتی که پیش‌تر گرفته شده بود بدست آمدند. در مثال مدل زبانی، اینجا دقیقاً جائی است که اطلاعاتی که در مورد جنسیت قبلی داشتیم را دور می‌ریزیم و اطلاعات جدید را اضافه می‌کنیم.

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

<sup>۹۴</sup>Input gate

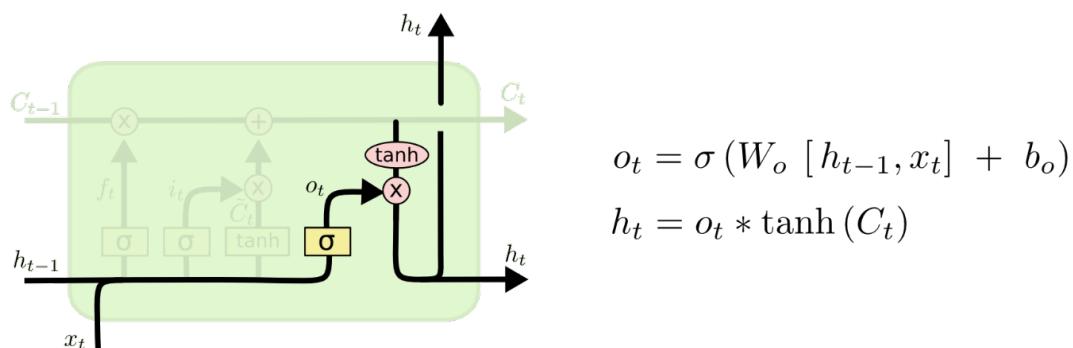


شکل ۲۶.۲: قدم دوم در اضافه کردن اطلاعات جدید به سلول حالت



شکل ۲۷.۲: بهروز رسانی اطلاعات در سلول حالت

می‌گیرد 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 تانژانت هایپربولیک (تا مقادیر بین  $-1$  و  $+1$  باشند) می‌دهیم و مقدار آن را در خروجی لایه سیگموید قبلی ضرب می‌کنیم تا تنها بخش‌هایی که مدنظرمان 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 مازول LSTM

## ۱۲.۲ یادگیری تقویتی

### ۱.۱۲.۲ مقدمه و بیشینه تاریخی

ادوارد ثورندایک<sup>۹۵</sup> پدر روانشناسی مدرن در سال ۱۸۷۴ میلادی در ایالت ماساچوست آمریکا متولد شد. وی در اوایل قرن ۲۰ میلادی آزمایشی انجام داد که باعث ارائه قانون اثر شد. او برای این آزمایش، گربه ای را در جعبه ای موسوم به جعبه معملاً 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 اشتباہی 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 پیشرفتی نایل آمد و در سال ۱۹۴۷ در سن ۷۴ سالگی، بدرود حیات گفت. در سال ۲۰۰۲ رتبه ای از برترین روانشناسان تاریخ ارائه شد که ثورندایک جزء ۱۰ روانشناس برتر تاریخ قرار گرفت. می‌توان مهمن ترین کشف وی را، اثبات وجود یادگیری تقویتی در روانشناسی دانست.

شاید ریچارد بلمن<sup>۹۶</sup> (مختصر الگوریتم بلمن-فورد) را بتوان اولین کسی دانست که یادگیری تقویتی را وارد هوش مصنوعی ساخت. در اوایل دهه ۱۹۵۰ بلمن مسئله ای 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 خطاب 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

<sup>95</sup>Edward Thorndike

<sup>96</sup>Richard E. Bellman

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 روش ساخت درخت تکاملی با استفاده از داده‌های توالی‌یابی تک سلولی

سرطان نامی است که به مجموعه‌ای از بیماری‌ها اطلاق می‌شود که از تکثیر مهار نشده سلول‌ها پدید می‌آیند. تحقیقات انجام شده نشان می‌دهد که سرطان در واقع یک فرآیند تکاملی از جهش‌های ژنتیکی، شامل حذف و تغییر تعداد کپی، حذف و تغییر تک نوکلئوتید‌ها، بازسازی و جایگزینی ژن‌ها در سلول‌های توموری است. در واقع تومور زمانی ایجاد می‌شود که یک سلول جهش یافته بتواند با عبور از سیستم دفاعی بدن زندگی کرده و تکثیر شود به گونه‌ای که نسبت مرگ به تولید آن گونه ایجاد شده بسیار کوچک تر از ۱ (۱) باشد.

با پیشرفت تومور، ناهنجاری‌های ژنتیکی مختلف منجر به افزایش گروه‌های جمعیتی ناهمگنی به نام کلون می‌شود. فرآیند تکاملی همه این کلون‌هارا می‌توان با یک درخت فیلوزنی و آنالیز فیلوزنیتیکی از چندین کلون سلولی سرطانی مدل‌سازی کرد که می‌تواند مطالعه انواع تومور را تسهیل کند. ساختار و الگو‌های درون این درخت میزان وابستگی بین گونه‌های خاص را با توجه به تعداد و فواصل بین اجداد مشترک-شان تعیین می‌کند. درخت‌های فیلوزنی عملکرد بارزی در توصیف فرآیند توسعه تومور دارند که بهتر از دیگر الگوریتم‌های مشابه عمل می‌کنند. تحلیل تپولوژی درخت درخت پیشرفت تومور نشان می‌دهد که مسیر توسعه تومور در طول مراحل مختلف تشکیل تومور، تا حد زیادی تغییر می‌کند و البته نتایج مبتنی بر درخت بهتر از نتایج داده‌های بدست آمده از طریق روش-های دیگر در تشخیص تومور می‌باشد. در حال حاضر ظهور تکنولوژی-های برپایه DNA یک سلول منفرد، با هدف افزایش دانش از جنبه‌های مختلف بیولوژی سرطان، شامل بررسی زیرساخت کلونال، ردیابی تکامل تومور، شناسایی زیرکلون-های نادر و درک ریزمحیط-های سرطانی در پیشرفت تومور، به یاری محققان این حوزه آمده و بالاترین وضوح را از تاریخچه سرطان (درخت فیلوزنی) فراهم کرده است. در واقع از آنجایی که در روش-توالی-یابی تک سلولی گونه‌های مختلف از ابتدا از هم جدا می‌شوند، از نقطه منظر از دست دادن نوع در زیرجمعیت بافت مورد آزمایش نداریم و به همین دلیل دقت این روش نیز نسبت به روش انبوه بالاتر می‌باشد. در کتاب مزايا اين روش، معايبي چون، هزينه بالا، از دست دادن سلول-ها، جهش ثانويه در هنگام كشت، از دست دادن ميزان فراوانی درون تومور حقيقي و زمان-گير بودن فرآيند نمونه گيري اشاره كرد. در ابتدا استفاده از روش-های توالي-یابي انبوه بدليل اينكه حجم بالايی از اطلاعات در اثر اين توالي-یابي ايجاد می‌شود، از محبوبیت بيشتری برخوردار بود اما با پیشرفت تکنولوژی و ظهور روش-های نوینی چون توالي-یابی تک-سلولی اين مهم دچار تغيير شد. در روش توالي-یابی انبوه، نمونه-برداری بر روی تعداد بسیار زيادي سلول (از محدوده‌ی هزار تا میليون سلول) صورت می‌گرفت و حجم بالاي داده‌ها و امكان تفکيك پايان نواحي ناهمگن، اطلاعات کافی از ساختار درون تومور و ناهمگنی-های درون توموري بدست نمي‌داد. در مقابل، در روش توالي-یابي تک-سلولی، اگر-چه ميزان هزينه نمونه-برداری افزایش قابل-توجهی داشت و يا ميزان اطلاعات از دست رفته و نويز موجود در داده-های توالي-يافته بالا بود.

## 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 روش‌های 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 دو روش پیشنهادی متفاوت را بیان خواهیم نمود. در روش اول که به رویکردهای پیشین نزدیک‌تر است 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<sup>۱</sup> به آن افزود و پاسخ گرفت. اما روش دوم کاملاً متفاوت بوده و با رویکردی جدید در حوزه یادگیری ماشین همراه است که به کمک یادگیری تقویتی به حل مسئله مورد نظر می‌پردازد.

#### 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<sup>۲</sup> هستند که حاوی اطلاعات جهش‌های ماتریس ژن-سلول (SNV) و اطلاعات مربوط به حذف و تغییر تعداد کپی هستند.

<sup>1</sup>Copy number variation (CNV)

<sup>2</sup>Ascii

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

### ۳.۴ روش پیشنهادی برای مدیریت داده‌های از دست رفته

در این بخش به معرفی روش پیشنهادی پرداخته خواهد شد. در ابتدا به دلیل وجود داده‌های از دست رفته در پایگاه داده‌های مورد استفاده به بررسی و رویکرد حل این مشکل خواهیم پرداخت و در ادامه پس از معرفی روش‌های پیشنهادی برای آن و هر کدام از آن‌ها را به طور مفصل شرح خواهیم داد. 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 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 روشی پیشنهادی بررسی خواهد شد.

### ۱.۳.۴ 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 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 خطای 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 جایگذاری 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

آنده 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  $= b(i)$  باشد، در این صورت  $(i)$  حتماً باید باشد و گرنه فرض اولیه مدل مکان‌های بینهایت نقض خواهد شد. اما اگر  $= 1(b)$  باشد، آنگاه نتیجه خاصی نمی‌توان گرفت و باید به دنبال نمونه والد  $a$  یعنی نمونه  $d$  باشیم. حال اگر  $= 1(d(i))$  باشد، آنگاه  $(i)$  حتماً باید باشد. اما اگر  $= d(i)$  بود آنگاه انتخاب هر مقداری برای  $(i)$  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 خطایی باشند 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 پارمترهای خطایی که در نظر می‌گیریم و فاصله‌زنی ای که دارند می‌توانند در تصمیم‌گیری تاثیرگذار 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 می‌کیم که به نوعی با توجه به ارزشی که به سلول‌های نواده شده از سلول  $j$  می‌دهد سعی دارد تا اطمینان  $\circ$  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 تاثیرگذاری خود می‌تواند در مورد جایگاه جهش  $i$  برای سلول  $j$  نظر دهد.

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

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.

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

مقدار یک عددی بین (۰، ۱) است که پارامتری در جهت میزان ارزش دهی به نوادگان با فواصل مختلف می باشد.

هرچه این عدد بزرگتر باشد به معنی کم ارزش تر شدن نوادگان با فواصل بیشتر است و بلافاصله.

به همین صورت برای اولاد سلول  $j$  نیز می توان 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 داده های دیگر از دست رفته در محاسبه توابع است که به دو صورت می توان 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 ۵٪

یا فراوانی نسبی آن جهش در محاسبات است که ما رویکرد اول را در این گزارش استفاده خواهیم کرد.

نکته دوم وجود خطأ در داده هاست. برای مدیریت این مشکل می توان با مدل سازی خطأ که به صورت فرمول ۴.۴ بیان می شود، 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 مدل سازی خطأ می توان روابط قبلی را مجددا به صورتی که در ادامه آمده است بازنویسی کرد.

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

که در این صورت تابع  $F_p$  و  $F_a$  نیز به صورت زیر همراه با مدل‌سازی خطاب 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 تابع  $\hat{F}$  محاسبه شوند، خود می‌توانند معیاری برای ارزیابی پایگاهداده در دسترس و احتمال درستی فرض مدل مکان‌های بی‌نهایت باشند.

#### ۱.۱.۳.۴ تصادفی

پر کردن کاملاً تصادفی میس‌ها

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	$\sigma$
ماتریس متناظر درخت	$X_T$
تعداد سلول‌های نمونه	$N$
تعداد جهش‌ها	$M$
نرخ خطای مثبت کاذب	$\alpha$
نرخ خطای منفی کاذب	$\beta$

## جدول ۲.۴: پارامترهای مدل ریاضی

زمان خدمت‌دهی به بیمار در مرحله $k$ ام	$t_{ik}$
زمان فاری خدمت‌دهی به بیمار در محله $k$ ام	$\tilde{t}_{ik}$
مقدار بدینانه (حداکثر) برای زمان خدمت‌دهی به بیمار در مرحله $k$ ام	$t_{ik}^p$
محتمل‌ترین مقدار برای زمان خدمت‌دهی به بیمار در مرحله $k$ ام	$t_{ik}^m$
مقدار خوشبینانه (حداقل) برای زمان خدمت‌دهی به بیمار در مرحله $k$ ام	$t_{ik}^o$

## جدول ۳.۴: متغیرهای مدل ریاضی

متغیر صفر-یک تخصیص بیمار به تخت/اتاق عمل	$X_{ild_k}$
زمان شروع خدمت‌دهی به بیمار	$S_{ild_k}$
متغیر صفر-یک توالی بیماران	$Y_{ijkl_k}$
متغیر صفر-یک تخصیص جراح به بیمار	$V_{ni}$

## ۴.۴ روش پیشنهادی اول (درخت‌بازی)

## 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 داده‌هایی که miss شده‌اند با روش‌های زیر می‌تواند برای ورود به مرحله بعد تخمین زده شود.

## 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 زن‌ها و بعد دوم  $N$  برابر سلول‌های نمونه‌برداری شده می‌باشد. در هر خانه  $d_{i,j}$  یک بردار داده قرار دارد که حاوی اطلاعات زن  $j$  در سلول  $i$  می‌باشد.

#### ۱.۱.۵ پایگاه داده مصنوعی<sup>۱</sup>

با توجه به این نکته که از درخت فیلوژنی حقیقی<sup>۲</sup> 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 بپردازیم. 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  $n$ ،<sup>۳</sup> ایجاد می‌کنیم که  $n$  تعداد زن‌ها (جهش‌ها) بوده و  $\zeta$  عددی در بازه  $(0, \infty)$  است که یک پارامتر کنترلی است که وظیفه اش کنترل کلی تعداد

<sup>1</sup>Synthetic Dataset

<sup>2</sup>Ground-truth Phylogeny Tree

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<sup>۳</sup>

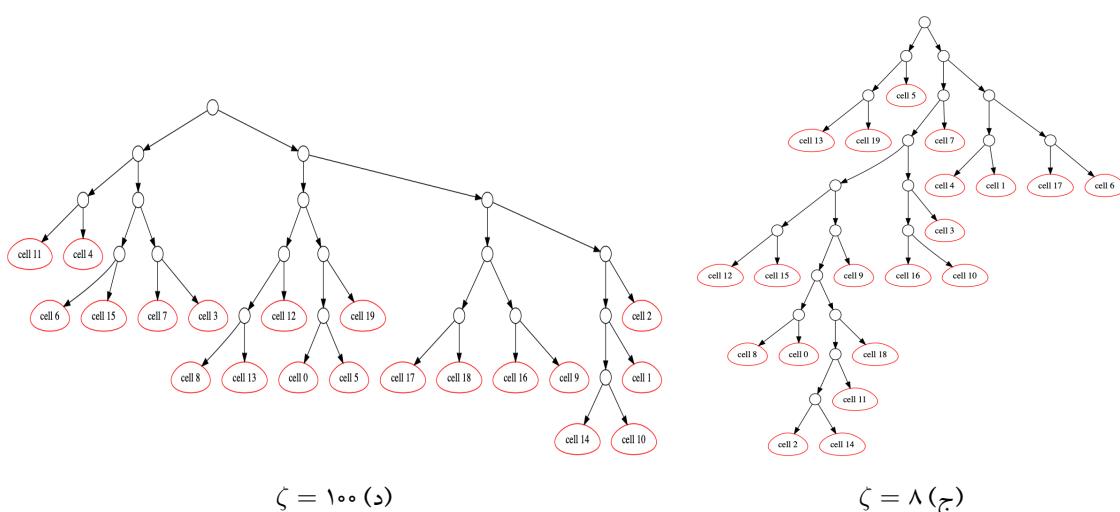
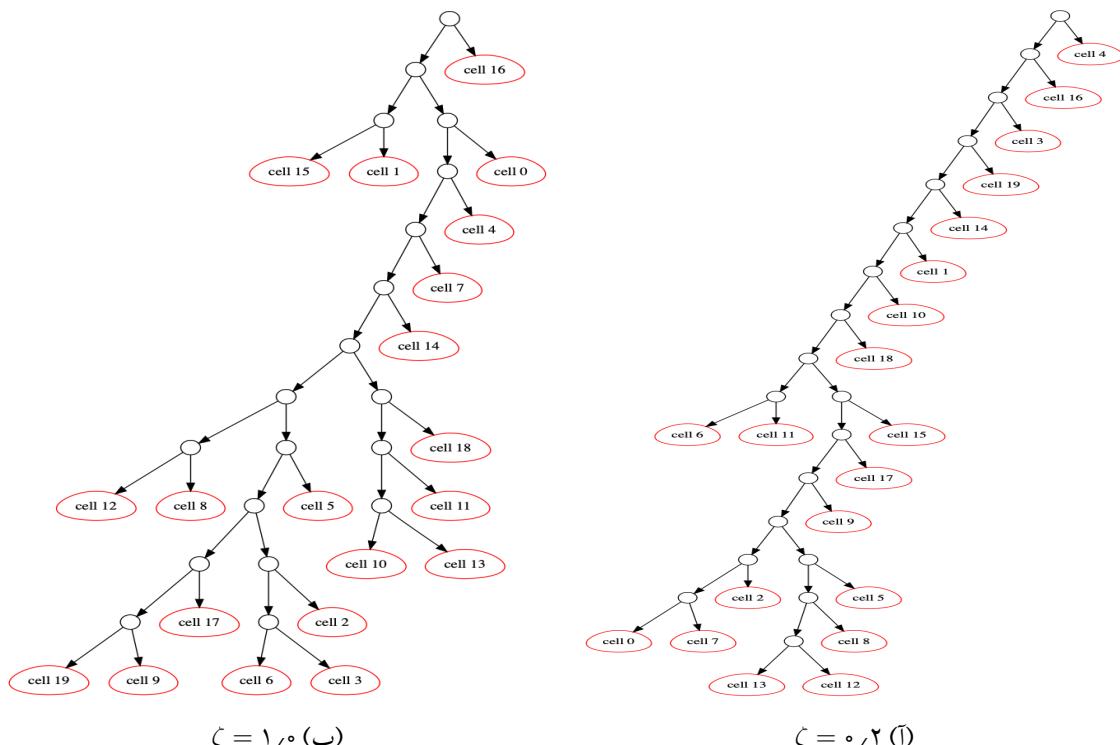
در این روش همان‌گونه که از نام آن مشخص است با استفاده از درخت تصادفی دودویی ژنولوژی به ساخت ماتریس داده ورودی مسلح می‌پردازیم که برای ساخت این دادگان از فرض‌های که در ادامه آمده است استفاده خواهیم کرد.

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  $1 = w_i$  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
- به جای دو گونه  $v, u$  یک گونه جدید  $uv$  با وزن  $w_{uv} = \frac{w_u + w_v}{\sqrt{\zeta}}$  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.

<sup>3</sup>Random Binary Genealogical Tree

پارامتر  $\zeta$  به گونه‌ای کنترل‌کننده میزان ناپایداری در طی نسل‌ها می‌باشد. بطوریکه نمونه‌ای از نتایج مقادیر مختلف آن برای  $n = 20$  در شکل ۱.۵ آورده شده است. پس از ساخت درخت تصادفی به سراغ مرحله بعد یعنی تبدیل درخت به ماتریس ژن-سلول  $E$  می‌رویم.



شکل ۱.۵: درخت فیلوزنی تصادفی تولید شده برای  $n = 20$  و  $\zeta$ ‌های مختلف

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.

#### روش دوم: با استفاده از درخت تصادفی جهش‌های <sup>۴</sup> ثانی

این روش نیز تا حدود زیادی مشابه روش قبل است با این تفاوت که در اینجا به جای اینکه درخت تصادفی را 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  $w_i = 1$  را اختصاص می‌دهیم که متناسب با روند حرکتی تومور به سمت آن جهش‌ها در مراحل بعدی خواهد بود.

<sup>4</sup>Random Mutation History Tree

- برای هر نود  $v$  تابع جرم احتمال را در ادامه به صورت  $F_v = \sum_{i=1}^{w_v} \frac{w_i}{w_v}$  بیان می‌شود در نظر می‌گیریم.
- با استفاده از  $F$  دو نود متمایز  $u, v$  را انتخاب می‌کنیم و به هم متصل می‌کنیم.
- به جای دو گونه  $u, v$  یک نود جدید  $uv$  با وزن  $w_{uv} = \sqrt{\zeta} \cdot \frac{w_u + w_v}{\zeta}$  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 ژن‌ها به هرکدام از این نودهای درخت است. برای این منظور به هرکدام از  $M$  نود یک ژن را به صورت تصادفی تخصیص می‌دهیم. پس از آن برای نهایی سازی درخت جهش‌ها از پارامتر دلخواه  $A = [\gamma * (M - 1) / \gamma]$  استفاده می‌کنیم که  $\gamma$  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.

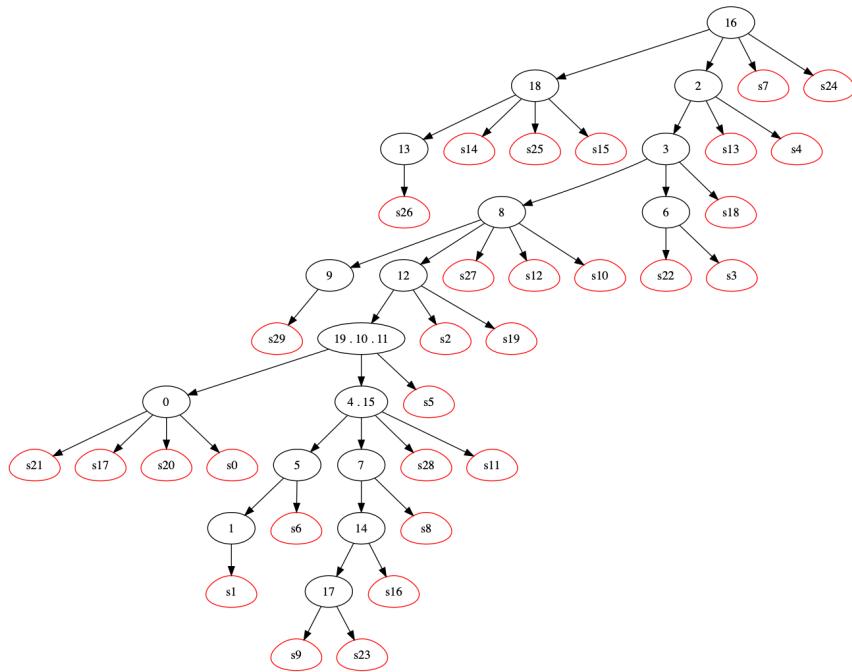
#### ۲.۱.۱.۵ تبدیل درخت به ماتریس ژن-سلول

با داشتن درخت (تولید شده با هرکدام از روش‌ها تفاوتی ندارد) در ادامه از فرض‌های مختلف در تولید ماتریس  $E$  می‌توان استفاده کرد.

##### فرض مدل مکان‌های بی‌نهایت<sup>۵</sup>

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

<sup>5</sup>Infinite Site Models



شکل ۲.۵: درخت جهش تصادفی با پارامترهای  $N = ۴۰, M = ۲۰, \zeta = ۱, \gamma = ۰.۱۵$

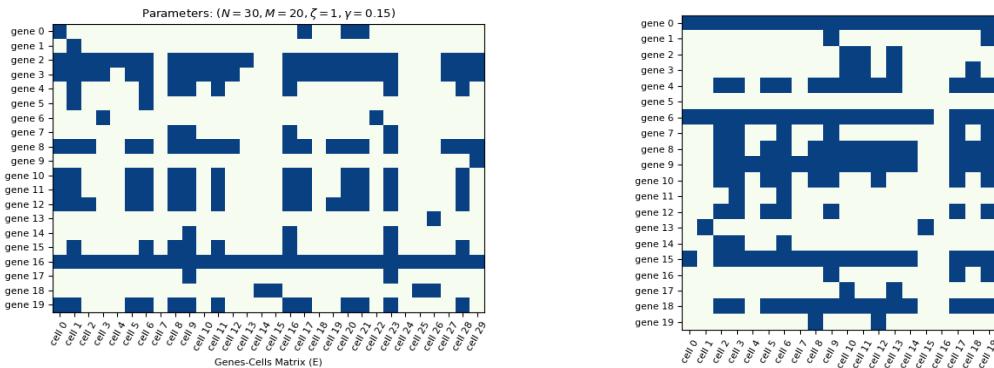
فیلوژنی کامل<sup>۶</sup> شناخته می‌شود.

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 آن که  $i$  شماره جهش و  $j$  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 خطای  $E$  را داریم که برای تصاویر دو روش درخت مرحله قبل در شکل ۳.۵ بدست آمده‌اند.

<sup>6</sup>Perfect Phylogeny Tree



(ب) ماتریس درخت شکل ۲.۵

(ا) ماتریس درخت شکل ۱.۵ ب.

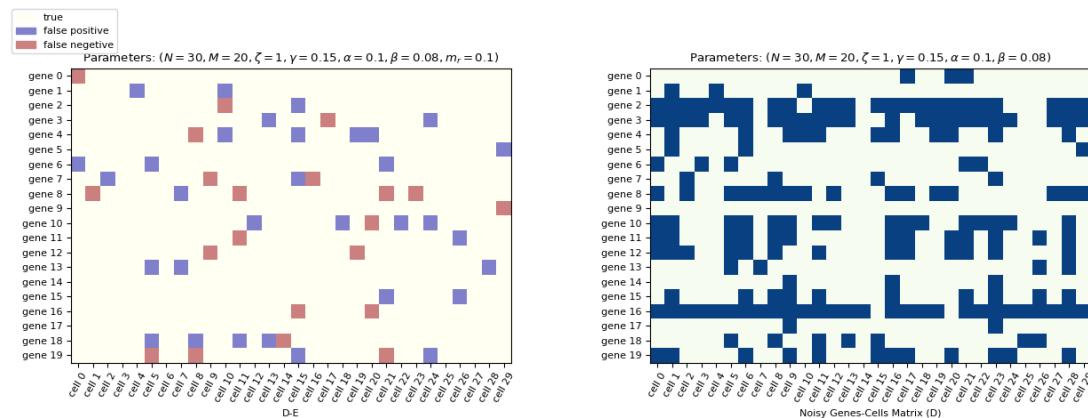
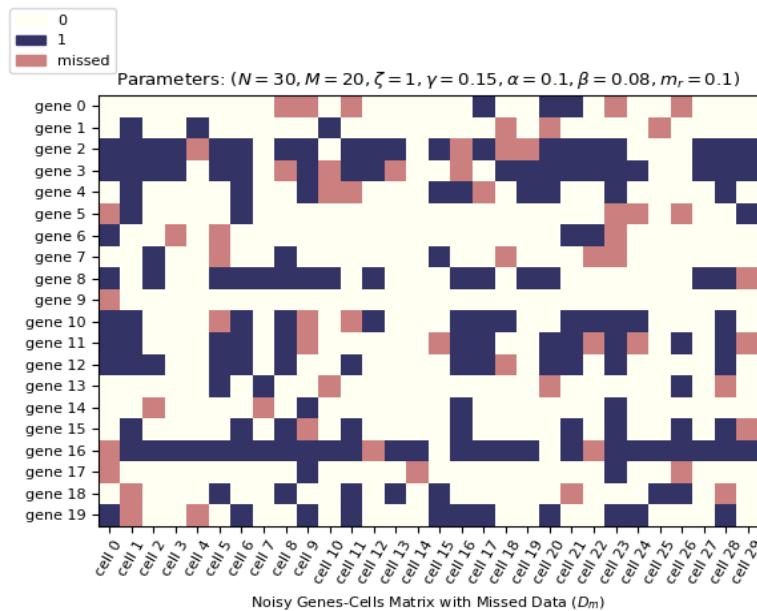
شکل ۳.۵: ماتریس‌های ژن-سلول ( $E$ ) بدست آمده از درخت‌های تصادفی ساخته شده

### 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  $\alpha = P(D_{ij} | E_{ij} = 0)$  و  $\beta = P(D_{ij} | E_{ij} = 1)$  است و همچنین  $m_r \in (0, 1)$  که نرخ داده‌های از دست رفته را مشخص می‌کند.

برای این منظور به ازای تمامی درایه‌های  $E$  هر بار یک عدد تصادفی با توزیع یکنواخت بین  $(0, 1]$  بوجود می‌آوریم و اگر عدد تولید شده کوچکتر از  $\alpha$  بود آنگاه ان درایه در ماتریس  $D$  را برابر با ۰ قرار می‌دهیم. به همین ترتیب مجدداً این بار برای درایه‌های  $E$  این کار را تکرار می‌کنیم و اگر عدد تصادفی تولید شده کوچکتر از  $\beta$  شد، درایه متناظر را در ماتریس  $D$  برابر با ۱ قرار می‌دهیم.

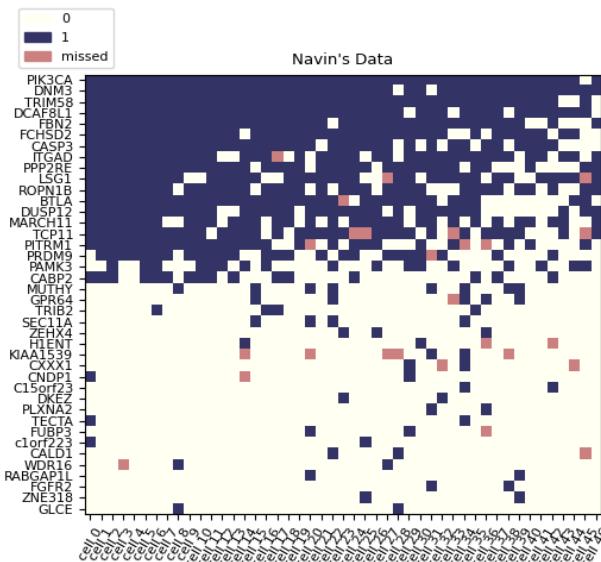
پس از اتمام کار نوبت به اضافه کردن داده‌های از دست رفته است. برای این منظور با نرخ  $m_r$  بعضی از درایه‌های ماتریس  $D$  را برابر با ۲ قرار می‌دهیم که به منزله در دسترس نبودن اطلاعات است. نام ماتریس نهایی را که شامل داده‌های از دست رفته است  $D_m$  می‌گزاریم. در ادامه تصاویر اضافه شدن نویز به ماتریس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

(ب) نویزی اضافه شده با پارامترهای  $\alpha = 0.1, \beta = 0.08$ (آ) ماتریس نویزی با  $\alpha = 0.1, \beta = 0.08$ (ج) ماتریس نویزی به همراه دادهای از دست رفته با پارامترهای  $\alpha = 0.1, \beta = 0.08, m_r = 0.1$ 

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<sup>۷</sup>

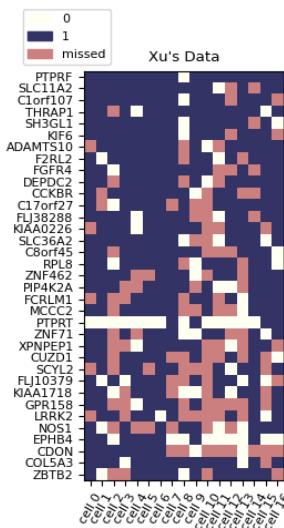
به عنوان پایگاه داده حقیقی از پایگاه داده استفاده شده در مقاله SCITE به عنوان پایگاه داده حقیقی اصلی استفاده خواهیم کرد که ماتریس داده ورودی آن به صورت شکل ۵.۵ می‌باشد. همچنین پایگاه داده حقیقی Xu



شکل ۵.۵: داده‌های حقیقی Navin در مقاله SCITE

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

<sup>7</sup>Real 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 فیلوژنی

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

بهینه برسیم.

### ۱.۱.۲.۵ استفاده از یک درخت نمونه نابهینه و تغییر اتصالات تا رسیدن به درخت فیلوژنی 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 جست‌وجوی حریصانه می‌باشد که طی شرایطی می‌توان انتظار داشت که به پاسخ بهینه دست یافته شود. این روش به نام روش زنجیره مارکو مونت-کارلو<sup>۸</sup> 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 می‌کند و ما به دنبال پایدارترین درخت هستیم که کمترین انرژی را داشته باشد. تعریف این تابع 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.

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

<sup>8</sup>Markov Chain Monte Carlo

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 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 برقرار 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 خطای 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  $\sigma_i$  برابر با بهترین نود (زن) برای اتصال نمونه بردار  $d_j$  است که بهترین جایگاه به صورت فرمول زیر انتخاب

می‌شود.

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

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.

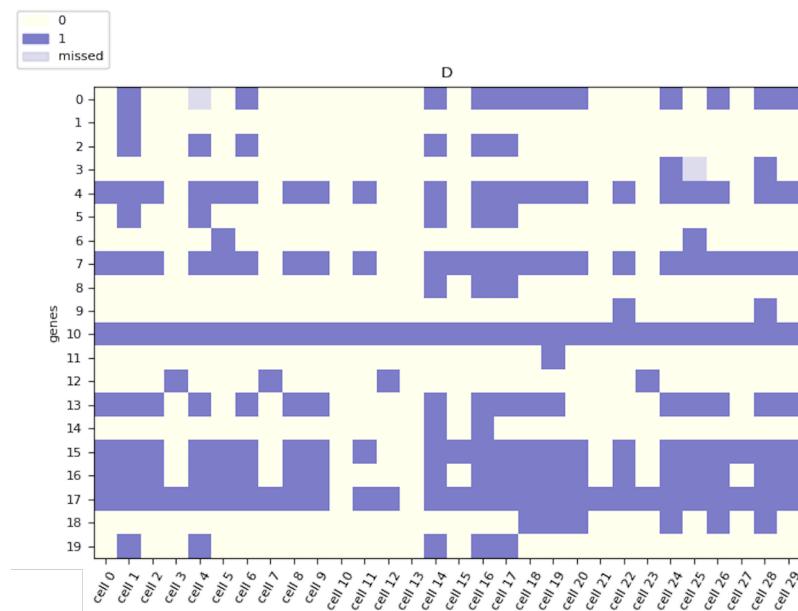
$$T_{op} = \min_{T \in \text{All possible } T} \left( \|D - \hat{E}_T\| \right) \quad (8.5)$$

## 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

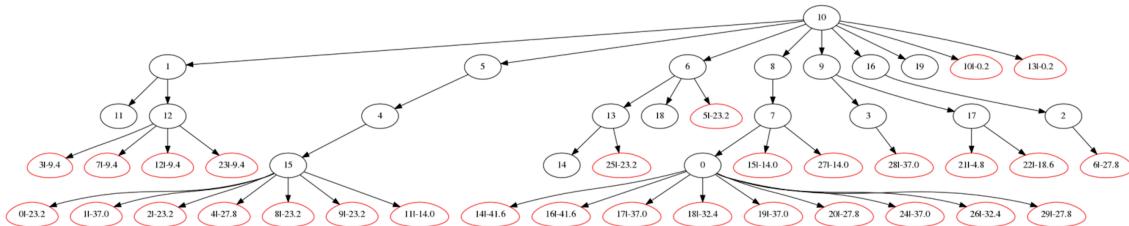
همان‌گونه که در بخش دوم توضیح داده شد با توجه به سختی دسترسی به پایگاه داده‌های حقیقی و اینکه در آن‌ها نیز حقیقت داده‌ها ( $E$ ) وجود ندارد تصمیم به ایجاد پایگاه داده‌ای مصنوعی گرفته شد که با کمک آن بتوان ارزیابی مناسبی از روش پیشنهادی و میزان کارایی و مقاومت روش را نسبت به تغییر پارامترها سنجید. فرض 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. پس از این مرحله اگر ۳۰۰۰ گام MCMC را اجرا نماییم می‌توانیم نتیجه حاصله را در شکل ۹.۵ مشاهده کیم. در

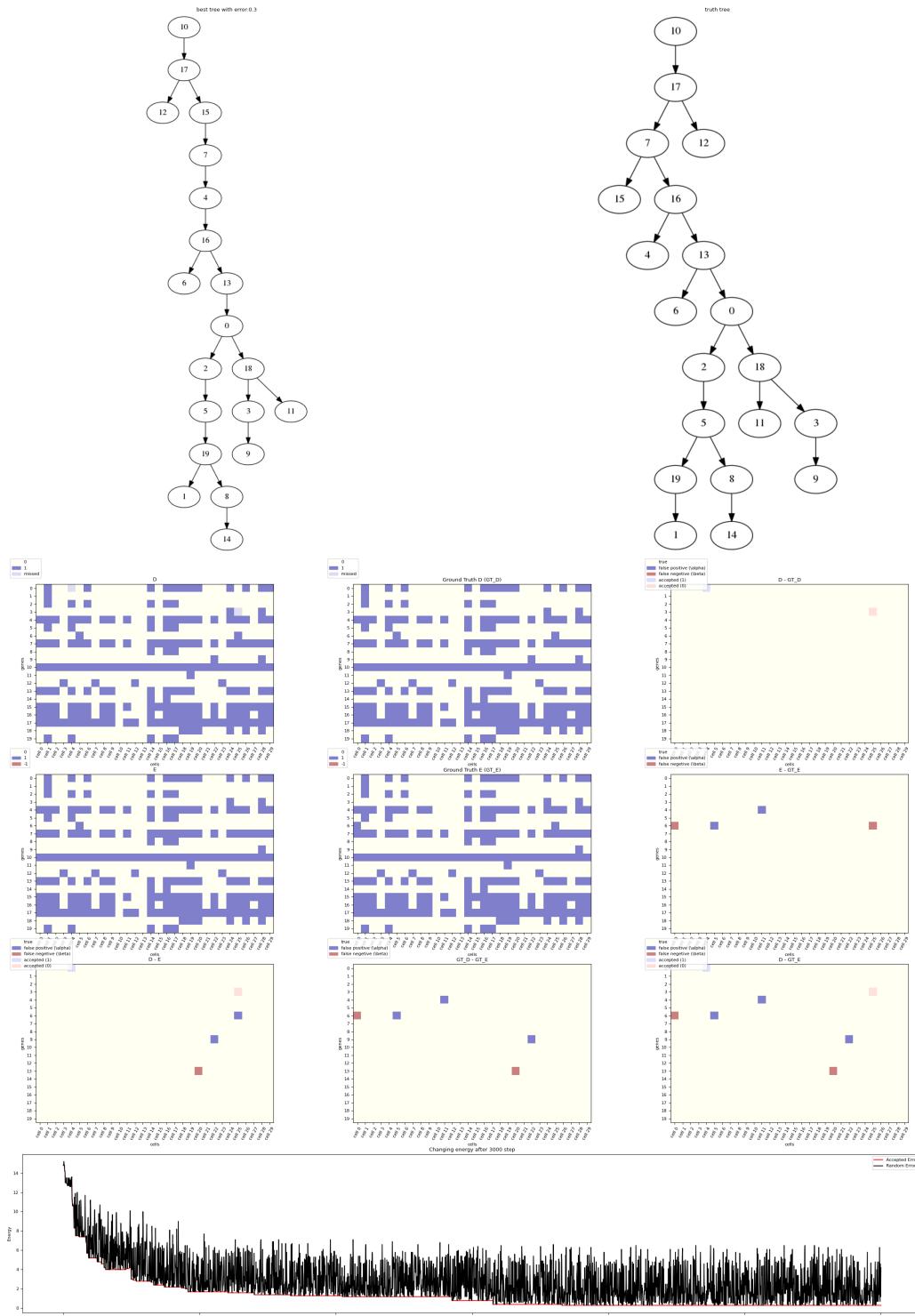
best tree with error:14.800000000000004



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

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 میان خطای میان 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.

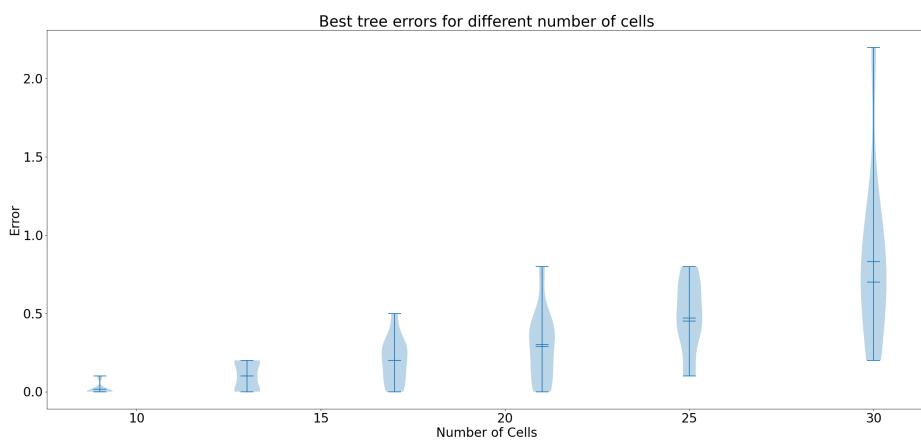
همان‌گونه که مشخص است در ماتریس  $D$  دوداده از دست رفته وجود دارد که یکی از آن‌ها در حقیقت جهش یافته و دیگر خیر. اگر ما در محاسبات خود این دوداده را در محاسبه خطای نظر نگیریم و با تغییر ۳ داده دیگر می‌توانیم به ماتریس  $\hat{E}$  (در شکل به نام  $E$  نوشته شده است) بررسیم که معادل بهترین درخت بدست آمده 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  $\frac{9916}{20*30} = 0.5$  بازسازی کند که عددی قابل قبول می‌باشد.



شکل ۹.۵: نتیجه اجرای روش پیشنهادی برای ماتریس شکل ۷.۵

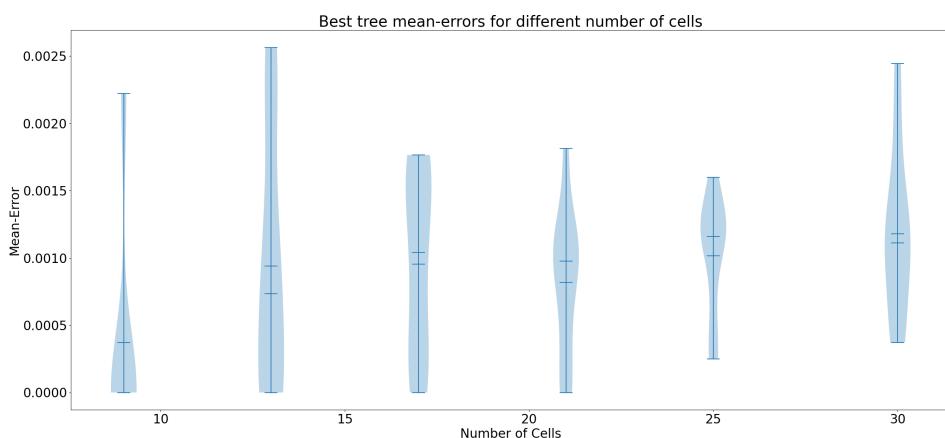
اما برای بررسی مناسب‌تر تعدادی تست را به ازای  $M$  و  $N$ ‌های مختلف اجرا نمودیم که به صورت خلاصه نتایج حاصل از آن در ادامه قابل مشاهده است.

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 خطا 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 خطا بخاطر ضعف روش پیشنهادی است یا ماهیت داده‌های ورودی میزان خطا را در هر اجرا بر تعداد خانه‌های ماتریس  $D$  تقسیم می‌کنیم که در آن صورت به نمودار شکل ۱۱.۵ می‌رسیم. در این نمودار جدید



شکل ۱۱.۵: نتیجه اجرای روش پیشنهادی برای تعداد نمونه‌های مختلف

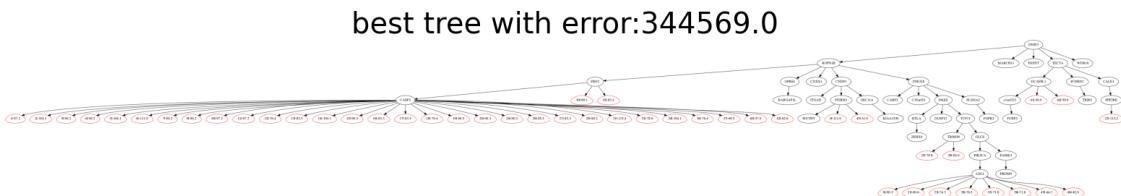
مشخص می‌شود که با افزایش اندازه ماتریس ورودی روش پیشنهادی سعی می‌کند تا خط را به ازای هر داده کنترل کند که نشان از کارآمدی روش پیشنهادی می‌باشد.

### 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.

#### ۱.۲.۳.۵ نتایج بهینه‌سازی درخت ثالث

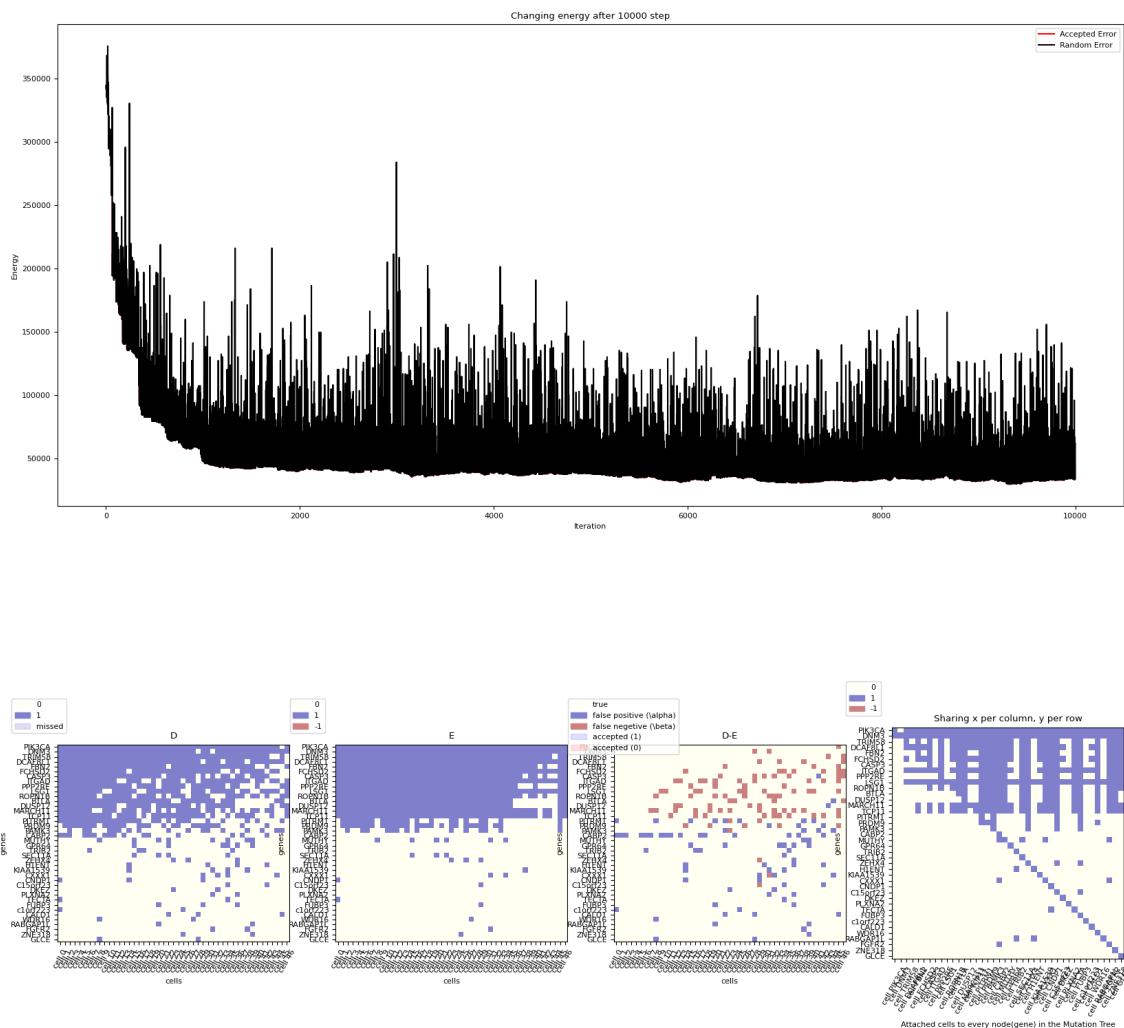
همان‌طور که در فصل قبل بیان شد، یکی از روش‌های بدست آوردن درخت فیلوزنی استفاده از یک درخت تصادفی نابهینه ثالثی بود که طی تکرار گام‌هایی سعی در تغییر اتصالات و یافتن درخت بهینه داشت که بتواند روند صحیح تغییرات ثالثی را در تومور مورد نظر نمایش دهد. نتیجه بدست آمده بر روی پایگاه داده حقیقی Navin به شرح 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 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.



شکل ۱۳.۵: نمودار تغییر انرژی در طی گام‌های مختلف

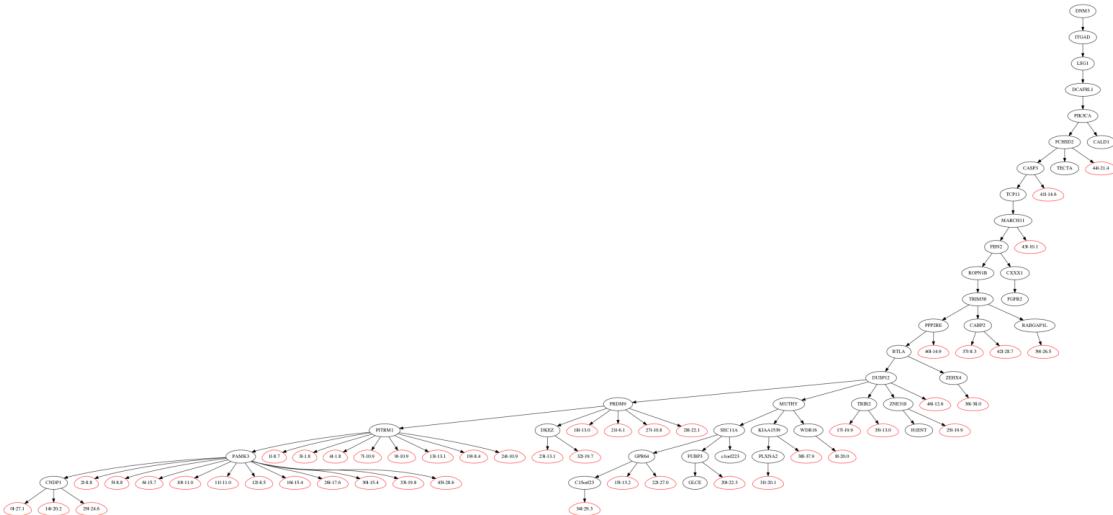
## ۴.۵ گام‌های آتی

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

best tree with error:29929.0



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

می شود تا شروع الگوریتم از نقطه بهتری باشد که در این صورت هم گام های لازم برای رسیدن به درخت بهینه می تواند کمتر شود و هم اینکه احتمال قرار گرفتن در نقاط اکسٹرمم نسبی را کاهش می دهیم.

## ۲.۴.۵ افزایش سرعت همگرایی MCMC

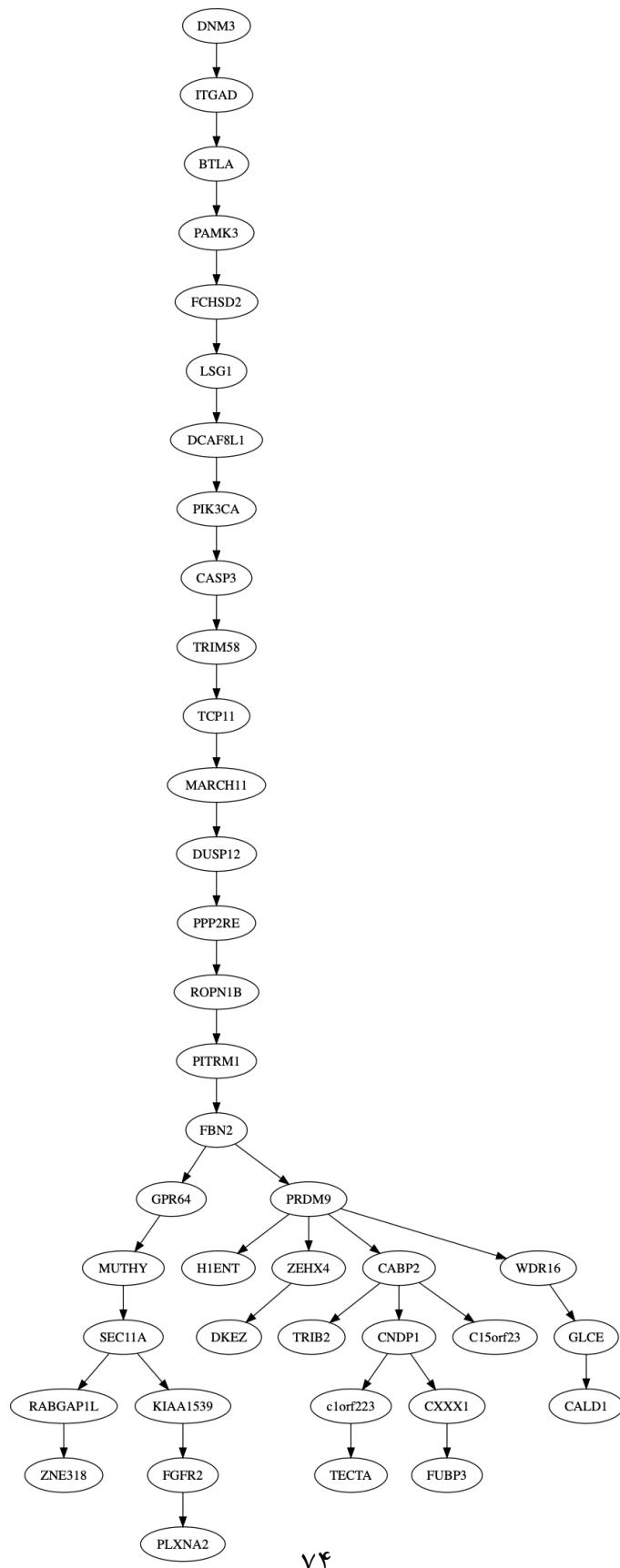
در این بخش نیاز است تا در دو قسمت روش پیشنهادی بهبود پاید.

## ۱.۲.۴.۵ تنویر 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 هر کدام از این روش‌ها در گام‌های مختلف احتمالاً بتوان سرعت همگرایی را افزایش داد.

#### ۲.۲.۴.۵ قرار دادن احتمال وزن دار به ازای هر انتخاب

در حال حاضر ما در هر کدام 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] 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.

- [10] 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.
- [11] Bishop, Christopher M. Pattern recognition. *Machine learning*, 128(9), 2006.
- [12] Burrell, Rebecca A and Swanton, Charles. Tumour heterogeneity and the evolution of polyclonal drug resistance. *Molecular oncology*, 8(6):1095–1111, 2014.
- [13] 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.
- [14] Cooper, Geoffrey M. *Elements of human cancer*. Jones & Bartlett Learning, 1992.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] Fearon, Eric R and Vogelstein, Bert. A genetic model for colorectal tumorigenesis. *cell*, 61(5):759–767, 1990.
- [20] 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.
- [21] Fisher, Rosie, Pusztai, Lazos, and Swanton, C. Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, 108(3):479–485, 2013.
- [22] Friedl, Peter and Wolf, Katarina. Plasticity of cell migration: a multiscale tuning model. *Journal of Cell Biology*, 188(1):11–19, 2010.

- [23] Fukushima, Kunihiro. Neocognitron. *Scholarpedia*, 2(1):1717, 2007.
- [24] Gelman, Andrew, Shirley, Kenneth, et al. Inference from simulations and monitoring convergence. *Handbook of markov chain monte carlo*, 6:163–174, 2011.
- [25] 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.
- [26] Greaves, Mel and Maley, Carlo C. Clonal evolution in cancer. *Nature*, 481(7381):306–313, 2012.
- [27] 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.
- [28] Hanahan, Douglas and Weinberg, Robert A. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.
- [29] Hanahan, Douglas and Weinberg, Robert A. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.
- [30] Handa, Osamu, Naito, Yuji, and Yoshikawa, Toshikazu. Redox biology and gastric carcinogenesis: the role of helicobacter pylori. *Redox Report*, 16(1):1–7, 2011.
- [31] Hastings, W Keith. Monte carlo sampling methods using markov chains and their applications. 1970.
- [32] 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.
- [33] LeCun, Yann, Bengio, Yoshua, and Hinton, Geoffrey. Deep learning. *nature*, 521(7553):436–444, 2015.
- [34] 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.

- [35] 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.
- [36] Nowell, Peter C. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.
- [37] 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.
- [38] 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.
- [39] Sokal, Alan. Monte carlo methods in statistical mechanics: foundations and new algorithms. In *Functional integration*, pages 131–192. Springer, 1997.
- [40] 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.
- [41] Stewart, BWKP and Wild, CP. World cancer report 2014. health, 2017.
- [42] Stratton, Michael R, Campbell, Peter J, and Futreal, P Andrew. The cancer genome. *Nature*, 458(7239):719–724, 2009.
- [43] 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.
- [44] 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.
- [45] Talbot, Simon J and Crawford, Dorothy H. Viruses and tumours—an update. *European Journal of Cancer*, 40(13):1998–2005, 2004.
- [46] Truninger, Kaspar, Menigatti, Mirco, Luz, Judith, Russell, Anna, Haider, Ritva, Gebbers, Jan-Olaf, Bannwart, Fridolin, Yurtsever, Hueseyin, Neuweiler, Joerg, Riehle, Hans-Martin, et al. Immunohistochemical analysis reveals high frequency of pms2 defects in colorectal cancer. *Gastroenterology*, 128(5):1160–1171, 2005.

- [47] 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.
- [48] 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.
- [49] 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.



# واژه‌نامهٔ فارسی به انگلیسی

	<b>default</b>
تشخیص به کمک کامپیوتر . . . . .	
diagnosis	
Single variant . . . . .	آدنین . . . . .
تغییر ساختاری . . . . .	آشکارسازی . . . . .
Internal covariate shift . . . . .	آنتروپی متقاطع دودویی . . . . .
Somatic single nucleotide variation	آسیب‌شناسی . . . . .
تغییرات بدnel تک نوکلئوتیدی . . . . .	اجماعی . . . . .
Copy number alteration . . . . .	استرومای اطراف . . . . .
Differentiation . . . . .	X-ray . . . . .
تمایز . . . . .	Concatenation . . . . .
Single cell sequencing . . . . .	انحراف استاندارد . . . . .
توالی‌بایی تک سلولی . . . . .	Stride . . . . .
Next generation sequencing . . . . .	اندازه‌گام . . . . .
توالی‌بایی نسل بعدی . . . . .	انکوژن . . . . .
Tumor . . . . .	Indel . . . . .
Potentially malignant tumor . . . . .	بایاس . . . . .
تومور بالقوه بدخیم . . . . .	برگشت . . . . .
Malignant tumor . . . . .	بعد . . . . .
Tumor Evolution . . . . .	بهینه‌سازی . . . . .
تکامل تومور . . . . .	Back-propagation . . . . .
Proliferative . . . . .	بیزی . . . . .
تکثیر . . . . .	Over-fitting . . . . .
Iteration . . . . .	Bayesian . . . . .
تکرار . . . . .	بیش‌پردازش . . . . .
Thymine . . . . .	Computer Vision . . . . .
تیمین . . . . .	تابع فعالیت . . . . .
somatic . . . . .	تابع هزینه . . . . .
جسمی . . . . .	تابع واحد اصلاح شده خطی (ReLU)
Single nucleotide mutation . . . . .	ترک آللی . . . . .
جehش تک نوکلئوتیدی . . . . .	Allelic dropout . . . . .
Driver mutation . . . . .	
جehش راننده . . . . .	
Single Somatic Mutation . . . . .	
جehش ساده بدnel . . . . .	
Passenger mutation . . . . .	
جehش مسافر . . . . .	
Point mutation . . . . .	
جehش نقطه‌ای . . . . .	
Long Short Term . . . . .	
حافظه‌ی کوتاه‌مدت بلند . . . . .	

Recurrent Neural Network	شبکه عصبی بازگشته	Memory (LSTM)
Intensity	شدت	حالت
Recognition	شناسایی	حذف تصادفی
Classification	طبقه بندی	حذف و تغییر تکنوکلئوتیدها variants (SNV)
Variant allele frequency	فرابواني تغییرات آلل	خال ملانوسیتیک
Activation	فعال سازی	خطا
Uterine fibroid	فیبرویید رحمی	خوانش
Segmentation	قطعه بندی	Overlapping read
Layer	لایه	داروهای سمیت سلولی
Zero padding	لایه گذاری صفر	درج
Mammogram	ماموگرام	دروازه
Metastases	متاستاز	دوره
False positive	مثبت کاذب	دی انای
Local	محلی	راديولژیست
Imaging modality	مدالیته تصویربرداری	روude بزرگ
False negative	منفی کاذب	روش ساخت درخت تکاملی
Antiparallel	موازی	evolutionary tree inference
Local receptive field	ناحیه ادراک محلی	رویکرد مبتنی بر داده
Heterogenetic	ناهمگن	Data driven
Tumor heterogeneity	ناهمگنی تومور	رگرسیون
Batch normalization	نرمال سازی دسته	Regularization
Biomarker	نشانگر زیستی	رگزایی
Feature map	نقشه ویژگی	زنجیره مارکوف مونت کارلو
Neuron	نورون	Monte Carlo (MCMC)
Nucleotid	نوكلئوتید	سرطان
Mapping	نگاشت	سرطان در محل
Hyperbolic tangent	هایپربولیک	سرطان سینه
Alignment	هم ترازی	Infiltrating immune cell
Artificial Intelligence (AI)	هوش مصنوعی	Cytosine
Network weight	وزن شبکه	Sigmoid
Read coverage	پوشش خوانش	Fully- connected feed forward neural network
Spontaneous	پیشرونده	شبکه عصبی کانولوشنی
Pixel	پیکسل	network

یادگیری تقویتی . . . . .	Reinforcement learning
یادگیری عمیق . . . . .	Deep learning
یادگیری ماشین . . . . .	Machine learning
حذف . . . . .	Deletion
مرگ . . . . .	Death

**T**

ژنوم جوانه‌زنی . . . . .	Germline genome
کانال . . . . .	Channel
کانولوشن . . . . .	Convolution
کرنل . . . . .	Kernel
کلونی . . . . .	Clonal
کواریانس . . . . .	Covariance
گلیوبالستوم . . . . .	Glioblastomas
گوانین . . . . .	Guanine



# واژه‌نامه انگلیسی به فارسی

Clonal .....	کلونی .....	A	فعال‌سازی .....
Colorectal carcinoma .....	روده بزرگ .....		تابع فعالیت .....
Computer Vision .....	بینایی کامپیوتر .....		آدنین .....
Computer-aided .....	تشخیص به کمک کامپیوتر .....		هم‌ترازی .....
diagnosis .....	diagnosis .....		ترک آللی .....
Concatenation .....	الحق .....		Angiogenic .....
Consensus .....	اجماعی .....		موازی .....
Convolution .....	کانولوشن .....		Artificial Intelligence (AI) .....
Convolutional neural .....	شبکه عصبی کانولوشنی .....		
network .....	network .....		
Copy number alteration .....	تغییرات شماره کپی .....	B	بک‌پروپگیشن .....
Cost function .....	تابع هزینه .....		نرمال‌سازی دسته .....
Covariance .....	کواریانس .....		بیزی .....
Cytosine .....	سیتوزین .....		Bias .....
Cytotoxic .....	داروهای سمیت سلولی .....		آنتروپی متقاطع دودویی .....
D			نشانگر زیستی .....
Data driven .....	رویکرد مبتنی بر داده .....		سرطان سینه .....
Death .....	مرگ .....	C	سرطان .....
Deep learning .....	یادگیری عمیق .....		سرطان در محل .....
Deletion .....	حذف .....		Channel .....
Detection .....	آشکارسازی .....		طبقه‌بندی .....
Differentiation .....	تمایز .....		
Dimension .....	بعد .....		
DNA .....	دی‌ان‌ای .....		
Driver mutation .....	جهش راننده .....		

Intensity .....	شدت .....	Dropout .....	حذف تصادفی .....
Internal covariate shift .....	تغییر همبستگی داخلی .....		
Iteration .....	تکرار .....		
		E	
K		epoch .....	دوره .....
Kernel .....	کرنل .....		
		F	
L		False negative .....	منفی کاذب .....
Layer .....	لایه .....	False positive .....	مثبت کاذب .....
Local .....	محلی .....	Feature map .....	نقشه ویژگی .....
Local receptive field .....	ناحیه ادراک محلی .....	Fully- .....	شبکه عصبی کاملاً متصل پیش خور .....
Long Short Term .....	حافظه کوتاه‌مدت بلند .....	connected feed forward neural network	
	Memory (LSTM)		
Loss .....	خطا .....	G	
M		Gate .....	دروازه .....
Machine learning .....	یادگیری ماشین .....	Germline genome .....	ژنوم جوانه‌زنی .....
Malignant tumor .....	تومور بدخیم .....	Glioblastomas .....	گلیوبالستوم .....
Mammogram .....	ماموگرام .....	Guanine .....	گوانین .....
Mapping .....	نگاشت .....	H	
Markov Chain .....	زنگیره مارکوف مونت کارلو .....	Heterogenetic .....	ناهمگن .....
	Monte Carlo (MCMC)	Hyperbolic tangent .....	هاپربولیک .....
Melanocytic nevi .....	حال ملانوسیتیک .....	I	
Metastases .....	متاستاز .....	Imaging modality .....	مدالیته تصویربرداری .....
N		Indel .....	ایندل .....
Network weight .....	وزن شبکه .....	Infiltrating immune cell .....	سلول ایمنی نفوذی .....
Neuron .....	نورون .....	Insertion .....	درج .....

Reinforcement learning .....	یادگیری تقویتی .....	Next generation sequencing .....	توالی‌یابی نسل بعدی .....
Replication .....	تکثیر .....		
reversion .....	برگشت .....	Nucleotid .....	نوکلئوتید .....

**S**

Segmentation .....	قطعه‌بندی .....
Sigmoid .....	سیگموید .....
Single cell sequencing .....	توالی‌یابی تک‌سلولی .....
Single nucleotide .....	جهش تک‌نوکلئوتیدی .....
	mutation
Single nucleotide .....	حذف و تغییر تک‌نوکلئوتیدها
	(SNV)
Single Somatic Mutation .....	جهش ساده بدنی .....
Single variant .....	تغییر ساختاری .....
somatic .....	جسمی .....
Somatic single .....	تغییرات بدنی تک‌نوکلئوتیدی .....
	nucleotide variation
Spontaneous .....	پیش‌روند .....
Standard deviation .....	انحراف استاندارد .....
State .....	حالت .....
Stride .....	اندازه گام .....
Surrounding stroma .....	استرومای اطراف .....

**T**

Thymine .....	تیمین .....
Tumor .....	تومور .....
Tumor Evolution .....	تکامل تومور .....
Tumor .....	روش ساخت درخت تکاملی .....
	evolutionary tree inference
Tumor heterogeneity .....	ناهمگنی تومور .....

**O**

Oncogene .....	انکوژن .....
Optimization .....	بهینه‌سازی .....
Over-fitting .....	بیش‌پردازش .....
Overlapping read .....	خوانش همپوشانی .....

**P**

Passenger mutation .....	جهش مسافر .....
Pathology .....	آسیب‌شناسی .....
Pixel .....	پیکسل .....
Point mutation .....	جهش نقطه‌ای .....
Potentially malignant tumor .....	تومور بالقوه بدخیم tumor .....
Proliferative .....	تکثیر .....

**R**

Radiologist .....	راديولوژیست .....
Read .....	خوانش .....
Read coverage .....	پوشش خوانش .....
Recognition .....	شناسایی .....
Rectified linear unit (ReLU) .....	تابع واحد اصلاح شده خطی (ReLU) .....
Recurrent Neural Network .....	شبکه عصبی بازگشته .....
Regression .....	رگرسیون .....
Regularization .....	رگولاتریزیشن .....

**U**

فیبرویید رحمی ..... Uterine fibroid .....

**V**

فراوانی تغییرات آلل .. Variant allele frequency ..

**X**

اشعه ایکس ..... X-ray .....

**Z**

لایه‌گذاری صفر ..... Zero padding .....

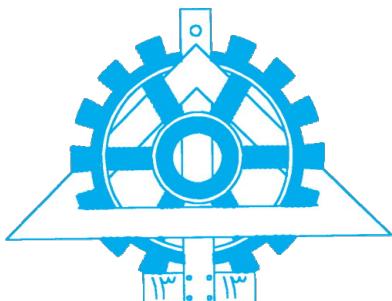


## **Abstract**

This thesis studies on writing projects, theses and dissertations using tehran-thesis class. It ...

**Keywords** SNV, CNV, Phylogenetic, Tree, Q-learning, Deep learning





University of Tehran  
College of Engineering  
**Faculty of New Science and  
Technology  
Network**



# **Inference of Phylogenetic Tree for Inter Tumor using Single Cell Mutations and CNV**

A Thesis submitted to the Graduate Studies Office  
In partial fulfillment of the requirements for  
The degree of Master of Science  
in Information Technology - Network Science

By:

**Afshin Bozorgpour**

Supervisors:

**Dr. Saman Haratizadeh and Dr. Abolfazl Motahari**

Jul 2021