
بسم الله الرحمن الرحیم

دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد مهندسی هوش مصنوعی

مدلی برای شناسایی و حذف نویز و بایاس از داده‌های Hi-C

نگارنده: سامان خاک‌مردان

اساتید راهنما

دکتر محسن رضوانی و دکتر علی‌اکبر پویان

اساتید مشاور

دکتر منصور فاتح و دکتر حمید علی نژاد رکنی

بهمن ماه ۱۳۹۷

در این صفحه صورت جلسه دفاع را قرار دهید. لازم است پس از صحافی این صفحه مجدداً توسط دانشکده مهر گردد و استاد راهنما با امضای خود اصلاحات پایان‌نامه را تایید کند.

تقدیم اثر

ماحصل آموخته‌هایم را با عشق تقدیم می‌کنم به:

پدر و مادر عزیزم که بودنشان تاج افتخاری بر سرم و نامشان دلیلی است بربودنم.

و تقدیم به دوستداران علم و دانش.

تشکر و قدردانی

با سپاس فراوان از لطف خدای مهربان

با تشکر از اساتید بزرگوارم جناب آقای دکتر محسن رضوانی و جناب آقای دکتر علی‌اکبر پویان، اساتید راهنمای ارجمند و جناب آقای دکتر منصور فاتح و جناب آقای دکتر حمید علی نژاد رکنی، اساتید مشاور محترم که شایسته هر نوع سپاس، تجلیل و تکریم اند و صبورانه با ارائه رهنمودها، انتقادها و پیشنهادهایشان در تمامی مراحل اجرای پایان‌نامه مرا حمایت و تشویق نمودند.

از استادان محترمی که در دوران تحصیلی‌ام در دورهٔ کارشناسی ارشد جهت آموزش و ارتقای علمی بنده زحمت کشیده اند سپاسگزارم.

تعهدنامه

اینجانب سامان خاک‌مردان دانشجوی دوره کارشناسی ارشد رشته هوش‌مصنوعی دانشکده مهندسی کامپیوتر دانشگاه صنعتی شاهرود نویسنده پایان‌نامه مدلی برای شناسایی و حذف نویز و بایاس از داده‌های Hi-C تحت راهنمائی دکتر محسن رضوانی متعهد می‌شوم.

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزارها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود. استفاده از اطلاعات و نتایج موجود در پایان‌نامه بدون ذکر مرجع مجاز نمی‌باشد.

چکیده

امروزه تحلیل اطلاعات پزشکی می‌تواند درک ما را نسبت به ساختار بدن انسان و همچنین عوامل تاثیرگذار بر آن را بالا برد. از این رو، پژوهشگران در حوزه علوم پزشکی برای جلوگیری از بروز بیماری‌ها در تلاشند این عوامل را شناسایی کنند. یکی از این عوامل تاثیرگذار، ساختار و نحوه قرارگیری رشته‌های DNA کروموزوم‌ها در یک فضای سه بعدی است. ایده اصلی این موضوع از جایی مطرح شد که در یک رشته درهم پیچیده DNA، فعل و انفعال بین دو ناحیه DNA زمانی که از نظر مکان فضایی به هم نزدیک هستند، تاثیر بسیار زیادی بر عملکرد بدن دارد. لذا پروتکل‌های آزمایشگاهی مختلفی برای بدست آوردن اطلاعات ساختاری کروموزوم‌ها توسعه یافته‌اند. یکی از این پروتکل‌ها روش Hi-C است. این پروتکل آزمایشگاهی دارای خطاهای سیستماتیک و آزمایشگاهی زیادی است. در نتیجه برای استفاده از این اطلاعات و فهم بهتر عملکرد بدن، نیاز به حذف نویز و استخراج اطلاعات معنی دار از این داده‌ها ضروری است. تا کنون پژوهش‌های زیادی در این زمینه انجام گرفته و همچنین روش‌های آماری مختلفی برای حل این مساله توسعه یافته‌اند. عموم این روش‌ها از یک رویکرد آماری سراسری برای تشخیص نویز استفاده کرده‌اند. حال آنکه با در نظر گرفتن یک رویکرد محلی می‌توان فرایند تشخیص نویز را بهبود داد. بنابراین در این رساله، دو رویکرد محلی و سراسری با هم ادغام شده‌اند. برای این منظور روشی بر پایه مدلی از شبکه عصبی عمیق به نام اتوانکدر، به دلیل توانایی این شبکه در حذف نویز، ارائه شده است. روش پیشنهادی ابتدا داده‌های Hi-C را به صورت آماری مدل می‌کند. سپس این روش با استفاده از شبکه عصبی مدل آماری بدست آمده را با توجه به تاثیر داده‌ها برروی یکدیگر بهبود می‌بخشد و به عبارت دیگر مدلی جدید را با استفاده از شبکه عصبی ایجاد می‌کند.

نتایج شبیه سازی نشان می‌دهد که در روش پیشنهادی با ۳۷۷۱ فعل و انفعال معتبر شناسایی شده نسبت به روش مرجع (GOTHiC) با ۳۰۲۷ فعل و انفعال معتبر شناسایی شده، عملکرد بهتری داشته

است. همچنین ضریب همبستگی بین مولفه فاصله و تعداد فعل و انفعالات بین دو ناحیه در روش پیشنهادی و روش مرجع به ترتیب برابر با ۰/۱۹۵۱- و ۰/۰۳۱۰- بوده است که در نتیجه روش پیشنهادی به فرض وابسته بودن تعداد فعل و انفعالات بین دو ناحیه به فاصله آن دو ناحیه، پایبندی بیشتری داشته است. به طور کلی در این پژوهش، ما نشان داده‌ایم که می‌توان با استفاده از شبکه عصبی به خوبی داده‌های Hi-C را مدل کرد و بر اساس مدل ایجاد شده توسط شبکه عصبی، نویز را حذف نمود و همچنین فعل و انفعالات معنی دار را در داده‌های Hi-C شناسایی کرد.

علاوه بر این، همراه با توسعه روش پیشنهادی، ابزاری به نام MHiC را برای حذف نویز و شناسایی فعل و انفعالات معنی‌دار در محیط R توسعه داده‌ایم. در این ابزار علاوه بر روش پیشنهادی، روش‌های GOTHiC، HiCNorm و FitHiC را نیز برای حذف نویز و شناسایی فعل و انفعالات معنی‌دار پیاده‌سازی نموده‌ایم. هر یک از روش‌های پیاده‌سازی شده در این ابزار بر خلاف پیاده‌سازی‌های موجود، توانایی دریافت ورودی از منابع HiC-Pro، HOMER، HiCUP و همچنین ورودی طراحی شده برای روش HiCNorm را دارا می‌باشند. همچنین ابزار مذکور داده‌های Hi-C را با استفاده از Contact map diagram و Arc diagram نمایش می‌دهد. ابزار ارائه شده امکان بصری‌سازی داده‌های Hi-C را فراهم نموده و مجموعه روش‌های موجود برای حذف نویز در این داده‌ها را یکجا ارائه می‌نماید.

کلمات کلیدی: Hi-C، شبکه عصبی عمیق، حذف نویز، بیوانفورماتیک، رگرسیون

فهرست مطالب

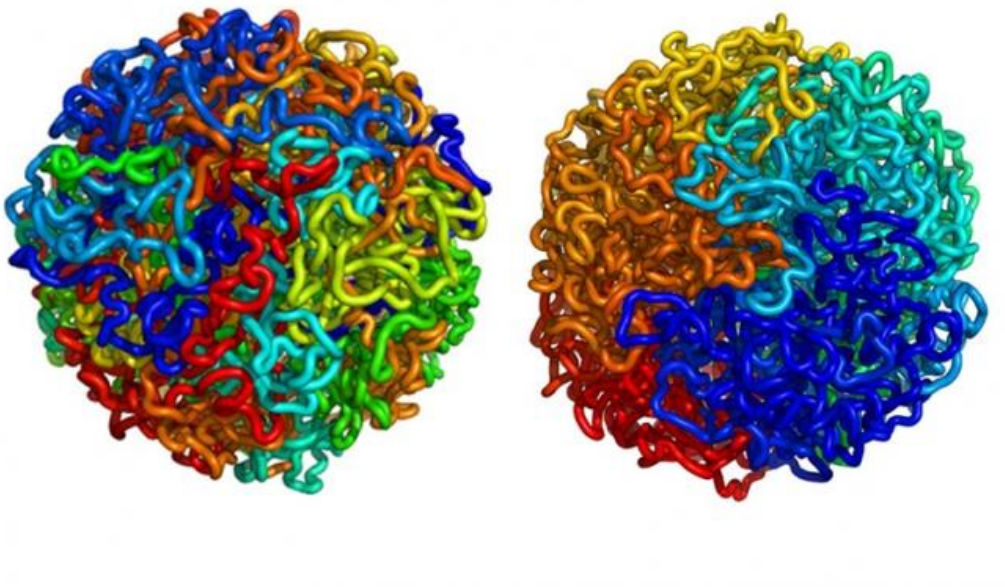
فهرست جداول

فهرست اشکال

فصل۱ : مقدمه

۱-۱ مقدمه

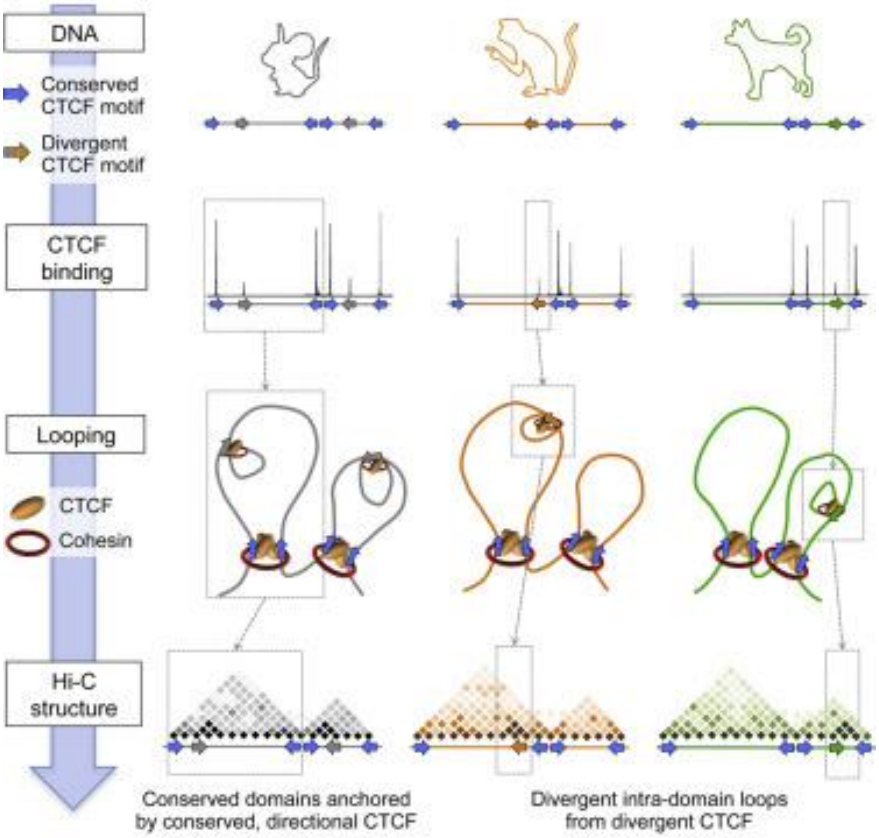
امروزه هوش مصنوعی در بسیاری از موضوعات پزشکی مانند پردازش تصاویر در کمک به پزشکان برای شناسایی یک بیماری خاص در یک بیمار، کاربرد داشته است [2] ,[1]. با این وجود مسائلی در حال حاضر وجود دارند که این مسائل خود به تازگی توسعه یافته‌اند. از این رو از روش‌های هوش‌مصنوعی در این مسائل به طور گسترده استفاده نشده است. برای نمونه می‌توان به موضوع ساختار و نحوه قرارگیری رشته‌های DNA کروموزوم‌ها در یک فضای سه بعدی اشاره کرد (شکل ۱-۱). درک این موضوع باعث افزایش دانش ما نسبت به نحوه عملکرد بدن، ساختار بدن، تاثیرات بیماری‌ها و همچنین عوامل تاثیر گذار بر بدن می‌شود [7]–[3]. به عنوان نمونه یکی از مسائلی که پژوهشگران به آن پرداخته‌اند، مساله ارتباط بین ویژگی‌های مختلف جانوران با ساختار کروموزومی آن‌ها بوده است. همان طور که در قسمت looping شکل ۲-۱ قابل مشاهده است که موجودات زنده مختلف دارای ساختار کروموزومی متفاوتی هستند و به طور کلی هر جانور دارای پیکربندی کروموزومی مخصوص به خود است.



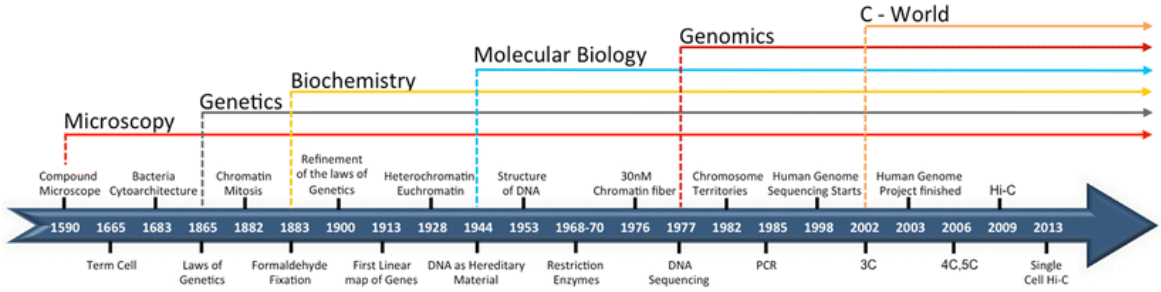
شکل ۱-۱ ساختار سه بعدی کروموزوم‌ها

تا به امروز تحقیقات زیادی برای گسترش دانش در مورد عملکرد بدن انجام شده است که در شکل ۱-۳ خلاصه‌ای از تاریخچه و گام‌های اساسی این تحقیقات ارائه شده است. یکی از جدیدترین روش‌ها

در تحقیقات بیولوژی که در سال ۲۰۰۲ ارائه شد و تا به امروز در حال توسعه بوده است، شناسایی و بررسی پیکربندی فضایی کروموزوم‌ها می‌باشد. برای یافتن ساختار سه‌بعدی کروموزوم‌ها (اطلاعات ساختاری کروموزوم‌ها) پروتکل‌های آزمایشگاهی مختلفی توسعه یافته‌اند [8]. این پروتکل‌ها برای پیش‌بینی ساختار سه بعدی کروموزوم‌ها، با استفاده از روش‌های مولکولی ناحیه‌هایی کروموزومی که در فضای سه بعدی به هم نزدیک هستند را به عنوان نواحی دارای فعل و انفعالات تشخیص می‌دهند. لذا با استفاده از این اطلاعات می‌توان ساختار سه بعدی کروموزوم‌ها را پیش‌بینی نمود. لازم به ذکر است که در اینجا زمانی فعل و انفعال بین دو ناحیه کروموزومی شکل می‌گیرد که با استفاده از برخی مولکول‌ها بتوان این دو ناحیه را به هم متصل نمود. به عبارت دیگر هر فعل و انفعال همان پیوند بین دو ناحیه است که توسط یک مولکول شکل گرفته است (شکل ۴-۱).



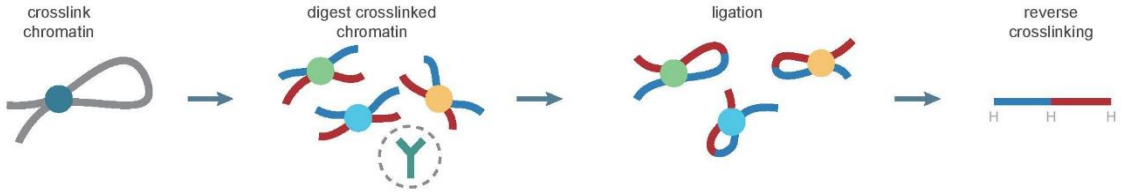
شکل ۲-۱ ارتباط بین ویژگی‌های جانداران مختلف با ساختار کروموزومی آنها [9]



شکل ۳-۱ گام‌های مهمی که به گسترش دانش کروموزوم‌ها منجر شد [10].

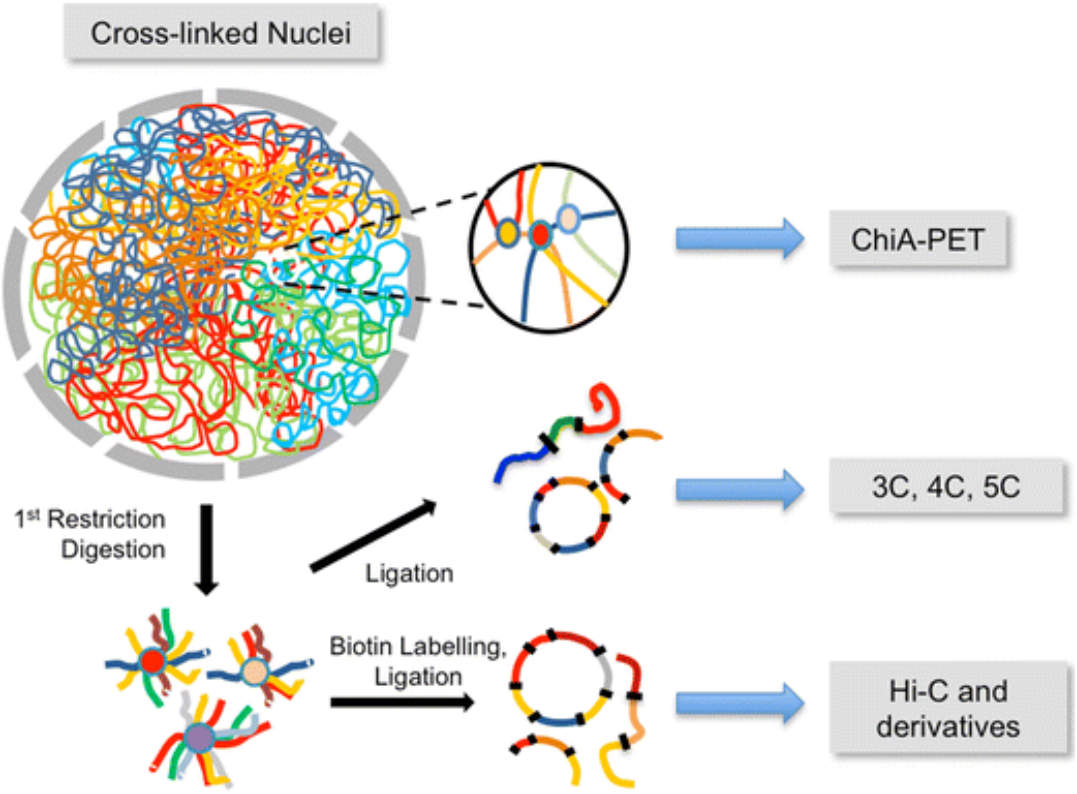
روش 3C یکی از اولین پروتکل‌هایی است که برای استخراج اطلاعات ساختاری کروموزوم‌ها توسعه یافته است [12] ,[11]. بسیاری از تکنیک‌های استخراج اطلاعات مانند 5C، 4C و همچنین پروتکل Hi-C از روش 3C مشتق شده‌اند [15]–[13]. همان طور که در قبل به آن اشاره شد، تمامی این پروتکل‌ها اطلاعاتی در مورد ساختارهای سه بعدی کرموزوم‌هایی که در سلول‌های زنده موجود هستند را فراهم می‌کنند. در واقع این مجموعه از تکنیک‌ها اطلاعات ساختاری کروموزوم‌ها (فعل و انفعالات داخل ساختار کروموزومی) را برای آنالیز و بدست آوردن اطلاعات کاربردی از این داده‌ها را فراهم می‌کنند. برای مشاهده نحوه پیدا کردن پیکربندی کروموزوم‌ها می‌توان به شکل ۲-۱ توجه کرد. اگر فرایند نمایش داده شده در این شکل را به صورت عکس بررسی کرد. مشاهده می‌شود که می‌توان پیکربندی کروموزم‌ها را بر اساس دادگان استخراج شده توسط روش‌ها مبتنی بر 3C (این داده‌ها در واقع اطلاعات خامی هستند که از کروموزوم‌ها استخراج شده‌اند.) پیش‌بینی کرد و لذا می‌توان ارتباط بین پیکربندی کروموزم‌ها و ویژگی‌های جانوران مختلف را مورد بررسی قرار داد.

لازم به ذکر است این روش‌ها به صورت کلی در مراحل استخراج اطلاعات با هم تفاوت دارند. به عنوان نمونه در شکل ۵-۱ مشاهده می‌شود در مرحله ligation روش Hi-C (از بیوتین در ایجاد پیوند استفاده می‌کند) با روش 3C، 4C و 5C تفاوت دارد. با وجود تفاوت در مراحل استخراج اطلاعات می‌توان گفت که تفاوت بنیادی بین این پروتکل‌ها، وسعت (میزان) استخراج اطلاعات پیکربندی کروموزموها است.



شکل ۴-۱ مراحل ایجاد فعل و انفعالات در پروتکل‌های مبتنی بر 3C [16]

در پروتکل Hi-C برخلاف پروتکل 3C (فعل و انفعالات بین دو ناحیه مورد بررسی قرار می‌گیرد.) تمامی فعل و انفعالات کروموزومی در سراسر ژنوم تعیین می‌شوند. این پروتکل نسبت به پروتکل‌های دیگر اطلاعات بیشتری را استخراج می‌کند. با این وجود این روش دارای خطا و نویز زیادی است . لذا برای بدست آوردن اطلاعات ساختاری صحیح، نیاز به حذف این خطاها می‌باشد.



شکل ۵-۱ نمایی از پروتکل‌های مبتنی بر پروتکل 3C [10]

تا کنون روش‌های مختلفی برای شناسایی خطاهای سیستماتیک و پیدا کردن اطلاعات معنی‌دار در داده‌های Hi-C انجام شده است. به عنوان نمونه می‌توان به روش‌های Fit-Hi-C، HiCNorm و GOTHiC اشاره کرد [17]–[19]. عموم روش‌های ارائه شده مانند سه روش فوق رویکرد آماری سراسری برای حذف نویز دارند. به عبارت دیگر این روش‌ها با استفاده از مدل‌های آماری و رگرسیون‌های مختلف

داده‌های Hi-C را با توجه به برخی مولفه‌های سراسری (مانند فاصله دو ناحیه تشکیل دهنده یک فعل و انفعال) مدل کرده و بر اساس مدل بدست آمده نویز را از داده‌های خام حذف می‌کنند.

قابل ذکر است که در نظر گرفتن یک رویکرد محلی همراه با رویکرد سراسری می‌توان فرایند تشخیص نویز را در داده‌های Hi-C بهبود ببخشد. لذا در این پایان‌نامه روشی با استفاده از شبکه عصبی عمیق به نام اتوانکدر برای تشخیص نویز در این داده‌ها توسعه داده‌ایم. در روش پیشنهادی برای ادغام دو رویکرد محلی و سراسری، از دو روش رگرسیون دوجمله‌ای منفی و شبکه عصبی عمیق اتوانکدر استفاده کرده‌ایم. همچنین پشتیبانی نکردن از ساختاره‌ای مختلف Hi-C مشکل دیگری هست که در پیاده‌سازی روش‌هایی مانند GOTHiC وجود دارد. بنابراین همراه با توسعه روش پیشنهادی، ابزاری به نام MHiC را برای حذف نویز و شناسایی فعل و انفعالات معنی‌دار در محیط R توسعه داده‌ایم. لازم به ذکر است که ابزار ارائه شده در کنار مجموعه روش‌های موجود برای حذف نویز امکان بصری‌سازی داده‌های Hi-C را نیز فراهم می‌کند.

۲-۱ شرح مساله

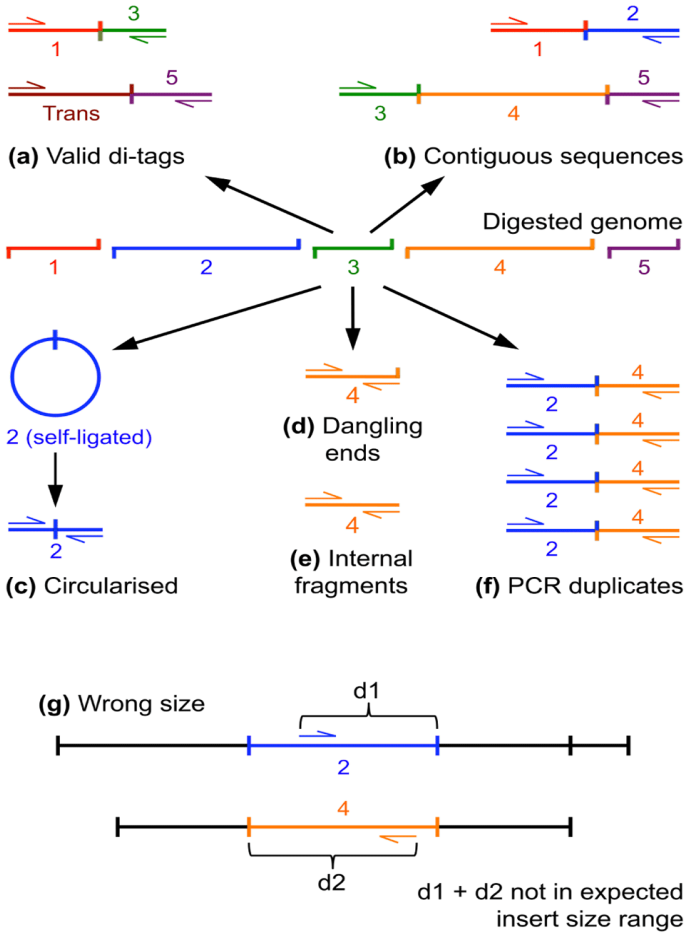
همان طور که در بخش قبل بیان شد، پروتکل Hi-C یک روش مبتنی بر 3C برای استخراج اطلاعات ساختاری کروموزوم‌ها می‌باشد. تاکنون پروتکل‌های مختلفی مانند in situ Hi-C نیز برای Hi-C ارائه شده است. این پروتکل‌ها در قسمت‌های مختلف استخراج اطلاعات مانند restriction enzyme با هم تفاوت دارند [20]. با این وجود تمامی این پروتکل‌ها اطلاعاتی را در مورد ساختارهای سه‌بعدی و فعل و انفعالات بین ناحیه‌های کرموزوم‌ها که در سلول‌های زنده موجود هستند را فراهم می‌کنند. پروتکل Hi-C تاثیر زیادی بر درک ما از سازمان (ساختار) ژنوم و همچنین تاثیر فعل و انفعالات کروموزومی داخل هسته سلول بر فعالیت‌های بیولوژیکی دارد. داده‌های حاصل از پروتکل‌های مختلف Hi-C به صورت مجموعه‌ای از اطلاعات برای شرح فعل و انفعالات کروموزوم‌ها می‌باشند. در فرایند استخراج اطلاعات پروتکل Hi-C، اطلاعات غیر ضروری و اشتباه استخراج می‌شوند. به عبارت دیگر اطلاعات استخراج شده

توسط پروتکل Hi-C دارای نویز و همچنین خطاهای سیستماتیک است. بنابراین برای استفاده از این داده‌ها و درک سازمان ژنوم، نیاز به شناسایی و حذف خطاهای موجود در این اطلاعات می‌باشد.

خطاهای موجود در اطلاعات Hi-C از منابع مختلفی حاصل می‌شوند. به طور کلی منابع تولید کننده خطا به دو دسته شناخته شده و ناشناخته تقسیم می‌شوند. یکی از منابع شناخته شده ایجاد خطا در دادگان Hi-C مرحله توالی‌یابی است که خطاهای حاصل از این مرحله را به عنوان خطاهای سیستماتیک (بایاس) در نظر گرفته می‌شوند. همچنین در دادگان Hi-C خطاهایی وجود دارند که منابع تولید این خطاها ناشناخته هستند و عموما به صورت تصادفی رخ می‌دهند. این دسته از خطاها به عنوان خطاهای تصادفی (نویز) شناخته می‌شوند. به عنوان نمونه با توجه به شکل ۶-۱ می‌توان مشاهده کرد که فعل و انفعالات غیر معتبر یا خطا می‌توانند به شکل‌های مختلفی در داده‌های Hi-C موجود باشند. در شکل ۶-۱ شش نمونه از خطاهای موجود در داده‌های Hi-C قابل مشاهده است. با توجه به این شکل، فعل و انفعالات غیر معتبر شامل، فعل و انفعالات همجوار[1]، فعل و انفعالاتی با یک ناحیه کروموزومی و فعل و انفعالات تکراری هستند. لازم به ذکر است که این خطاها خطاهایی هستن که منابع شناخته شده دارند و به راحتی در مرحله پیش پردازش قابل شناسایی هستند.

از نظر مفهومی می‌توان خطاهای سیستماتیک موجود در داده‌ها را به دو دسته ساده و پیچیده تقسیم کرد. منظور از خطاهای سیستماتیک و نویز ساده خطاهایی است که به راحتی در مراحل پیش پردازش شناسایی و حذف می‌شوند. به عنوان مثال self-ligations به معنای آن است که دو انتهای یک قطعه DNA با هم متصل شده‌اند. شناسایی این نوع از فعل و انفعالات تنها با بررسی ناحیه اول و دوم یک فعل و انفعال انجام می‌گیرد. در بسیاری از موارد خطاهای سیستماتیک و نویز در داده‌ها پنهان شده‌اند. به این نوع خطاها، خطاهای سیستماتیک و نویز پیچیده گفته می‌شود. در این نوع خطاها ممکن است تنها اثر این خطاها بر برخی مولفه‌ها قابل مشاهده باشد. در نتیجه تشخیص تمامی فعل و انفعالات غیرضروری و اشتباه، عملا کاری بسیار دشوار و غیر ممکن می‌باشد.

[1] contiguous sequence



شکل ۶-۱ نمای کلی از مصنوعات تولید شده توسط پروتکل Hi-C [21]

با توجه به پیچیدگی خطاهای موجود در در ساختار Hi-C، استفاده از تکنیک‌های جدید هوش مصنوعی نظیر یادگیری عمیق می‌تواند به تشخیص دقیق‌تر این خطاها کمک نماید. از این رو روشی را بر پایه ترکیبی از تکنیک‌های شبکه عصبی و تکنیک‌های آماری برای شناسایی نویز و خطاهای سیستماتیک توسعه می‌دهیم. لذا در این پایان‌نامه به بررسی امکان استفاده از یادگیری عمیق برای تشخیص خطاهای پیچیده پرداخته شده است. علاوه بر این تا کنون ابزار یکپارچه‌ای برای تحلیل روش‌های موجود برای تشخیص خطا در Hi-C توسعه داده نشده است. بنابراین در این پایان‌نامه ابزاری توسعه خواهیم داد که این امکان برای پژوهشگران فراهم شود که بتوانند با تکنیک‌های مختلف خطاهای موجود در داده‌های Hi-C را تشخیص داده و تحلیل نمایند.

۱-۳ ضرورت و اهمیت انجام پژوهش

به دلیل اینکه که در پروتکل‌های استخراج اطلاعات ساختاری کرموزوم‌ها خطا وجود دارد، درک بهتر روابط بین ساختارها سه‌بعدی کرموزوم‌ها و بیماری‌ها نیازمند آن است که قبل از استفاده از اطلاعات بدست آمده به واسطه این پروتکل، خطاهای سیستماتیک موجود در این اطلاعات حذف شده و اطلاعات معنی دار از میان این اطلاعات استخراج شوند. به عبارت دیگر این اطلاعات برای تحلیل بیشتر و فهم عملکرد بدن توسط زیست شناسان، آماده می‌شوند. به طور کل اهمیت حذف خطا در داده‌های Hi-C به صورت زیر است:

- بهبود در شناسایی ارتباط بین ساختارهای کروموزومی و بیماری‌های خودایمنی و انواع سرطان.
- بهبود در شناسایی ارتباط بین ساختارهای کروموزومی و عملکرد بدن.
- بهبود در شناسایی ارتباط بین ساختارهای کروموزومی و سیستم ایمنی بدن.

۱-۴ اهداف

هدف این پژوهش ارائه روشی برای حذف خطاهای سیستماتیک و پیدا کردن اطلاعات معنی‌دار از اطلاعات خام بدست آمده از پروتکل Hi-C می‌باشد. همچنین در راستای این موضوع، توسعه یک ابزار قدرتمند برای پیدا کردن اطلاعات معنی‌دار که بتواند از انواع مختلفی از پروتکل‌های Hi-C پشتیبانی کند، جزو اهداف این پژوهش است. به طور کل اهداف این پژوهش به شرح زیر می‌باشد:

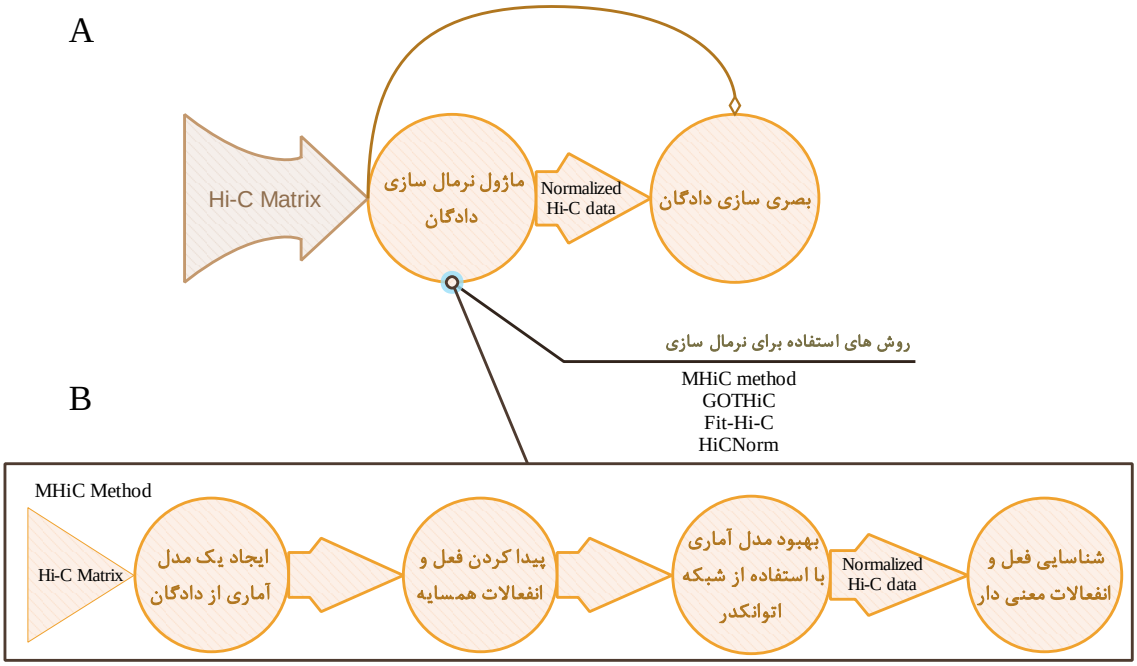
- ارائه روشی دقیق برای حذف خطاهای سیستماتیک از داده‌های Hi-C.
- توسعه روشی انعطاف‌پذیر که توانایی پردازش اطلاعات استخراج شده از پروتکل‌های مختلف Hi-C را داشته باشد.
- توسعه روشی که بتواند از تمامی مولفه‌های مهم و تاثیر گذار داده‌های Hi-C استفاده کند.

- توسعه ابزاری جامع برای پردازش داده‌های Hi-C.

۱-۵ مختصری از روش پیشنهادی

همان طور که در بخش‌های قبل به آن اشاره شد، روشی را برپایه ترکیبی از مدل‌های آماری و شبکه عصبی عمیق برای حذف نویز و شناسایی فعل و انفعالات معنی‌دار توسعه داده‌ایم. روش پیشنهادی برای حذف کردن نویز از داده‌های Hi-C، ابتدا این داده‌ها را به صورت آماری مدل می‌کند. برای مدل کردن این داده‌ها به به صورت آماری از مدل ارائه شده در روش GOTHiC استفاده نموده‌ایم. بعد از مدل کردن داده‌های Hi-C، مدل بدست آمده را با استفاده از شبکه عصبی اتوانکدر با توجه به تاثیرات فعل و انفعالات همسایه بر روی یکدیگر، بهبود می‌بخشد و به عبارت دیگر مدلی جدید را با استفاده از شبکه عصبی ایجاد می‌کند. به طور کلی روش پیشنهادی دارای پنج مرحله اساسی است. این مراحل مطابق با شکل ۷-۱ قسمت (B) به شرح زیر است:

۱. دریافت داده‌های Hi-C به صورت ماتریسی از فعل و انفعالات.
۲. بدست آوردن مدلی آماری ازداده‌های Hi-C با استفاده از مدل ارائه شده در روش GOTHiC.
۳. آماده‌سازی داده‌های Hi-C برای استفاده از شبکه عصبی.
   ۳/۱. مشخص کردن فعل و انفعالات همسایه با استفاده از روش خوشه بندی k-means.
   ۳/۲. نرمال‌سازی مقادیر فعل و انفعالات برای آموزش شبکه عصبی.
۴. مدل کردن داده‌های Hi-C با استفاده از شبکه عصبی اتوانکدر.
۵. پیدا کردن فعل و انفعالات معنی‌دار.



۶.

۷. **شکل ۷-۱ شمایی از ابزار MHiC همراه با روش پیشنهادی**

۶-۱ مختصری از ابزار توسعه داده شده

همان طور که در بخش‌های قبل به آن اشاره شد، ابزاری به نام MHiC را با ساختار خطلوله‌ای و ماژولار برای حذف نویز و شناسایی فعل و انفعالات معنی‌دار در محیط R توسعه داده‌ایم. ابزار ارائه شده، مجموعه روش‌های موجود برای حذف نویز در این داده‌ها را یکجا ارائه می‌نماید. همچنین این ابزار امکان بصری‌سازی داده‌های Hi-C را برای پژوهشگران فراهم نموده است. این ابزار را با هدف جمع‌آوری روش‌های مختلف حذف نویز در یک مجموعه توسعه داده‌ایم. مراحل انجام فرایند در ابزار ارائه شده مطابق با شکل ۷-۱ قسمت (A) به شرح زیر است:

۱. دریافت ماتریس داده Hi-C از کاربر توسط رابط کاربری گرافیکی یا فرمان R.

۲. پردازش داده‌های Hi-C و حذف نویز از این داده‌ها با استفاده از چندین روش پیاده‌سازی شده در این ابزار.

۳. بصری‌سازی داده‌های Hi-C.

۱-۷ فرضیات تحقیق

در بدنه اصلی تحقیق چندین فرض وجود دارد که از داده‌های واقعی گرفته شده است. بدیهی است بدون این فرضیات نمی‌توان اطلاعات معنی‌دار و همچنین خطاهای سیستماتیک را شناسایی کرد. برای اینکه بدانیم آیا فعل و انفعال بین دو ناحیه درست یا به عبارت دیگر معنی‌دار است، از تعداد فعل و انفعالات بین این دو ناحیه که به صورت آزمایشگاهی بدست آمده‌اند استفاده می‌کنیم. از این رو بیشتر فرضیات مساله روی بحث تعداد فعل و انفعالات بین نواحی مطرح شده است. این فرضیات باعث می‌شود که با دقت بیشتری این تعداد فعل و انفعالات را مدل کنیم و تخمینی از مقدار واقعی فعل و انفعالات بدست آوریم. برخی از فرض‌های اصلی مساله به شرح زیر می‌باشد [22] ,[18]:

- تعداد فعل و انفعالات بین دو ناحیه کروموزومی با فاصله بین این دو ناحیه نسبت عکس دارد.
- تعداد فعل و انفعالات بین دو ناحیه کروموزومی تابعی از معیار GC content می‌باشد.
- میزان نویز در دادگان نسبت به داده‌های صحیح کم است.

۱-۸ ساختار پایان‌نامه

این پایان‌نامه شامل پنج فصل می‌باشد که فصل یک به بیان کلیات تحقیق شامل مقدمه‌ای از موضوع مورد بحث و اهداف آن اختصاص دارد. در فصل دوم به مرور ادبیات و تشریح مسالهو مفاهیم اولیه بکار رفته در این تحقیق می‌پردازیم. فصل سوم را به روش پیشنهادی و تشریح کامل مراحل کار اختصاص داده‌ایم. در فصل چهارم به بررسی آزمایشات و ارزیابی نتایج می‌پردازیم و در نهایت و در فصل پنجم خلاصه‌ای از مطالب ارائه شده و سوی کارهای آینده ارائه خواهد شد.

فصل ۲ ادبیات و پیشینه تحقیق

۲-۱ پیش‌درآمدی بر ادبیات و پیشینه تحقیق

با توجه به ضـــرورت انجام تحقیق، تحقیقات مختلفی برای پردازش داده‌های Hi-C، به خصـــوص شناسایی خطاهای سیستماتیک و پیدا کردن اطلاعات معنی‌دار در این داده‌ها انجام شده است. برخی از این تحقیقات مانند GOTHiC با استفاده از تکنیک‌های نسبتا ساده آماری و برخی دیگر با استفاده از تکنیک‌های ریاضـــی پیچیده‌تر و اســـتفاده از معیارهای دیگر، خطاهای موجود در داده‌های Hi-C را شناسایی کرده و سپس حذف می‌کنند. هریک از روش‌های ارائه شده دارای معایبی مانند دقت کم در شناسایی خطاهای سیستماتیک، پشتیبانی نکردن این روش‌ها از انواع داده‌های Hi-C و حساس بودن به بعضی از پارامترها می‌باشند [23]. در این فصل ابتدا به صورت مختصر به بررسی برخی از مفاهیم استفاده شده در این بحث پرداخته می‌شود. سپس تحقیقات انجام شده در راستای این موضوع بررسی می‌شوند.

۲-۲ ادبیات و تعاریف مساله

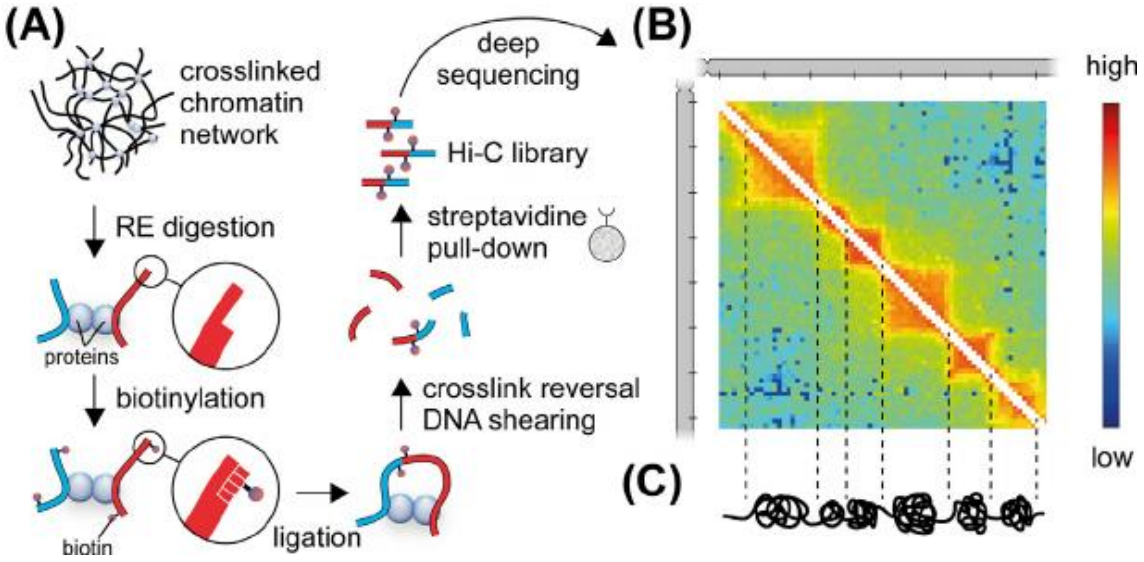
در این بخش شرح مختصری از مفاهیم و روش‌های مورد استفاده در روش پیشنهادی آورده شده است.

۲-۲-۱ داده‌های Hi-C و پردازش آن‌ها

همان‌طور که در فصل قبل به آن اشاره شد، پروتکل Hi-C اولین بار در سال ۲۰۰۹ بر مبنای روش 3C ارائه شد. این پروتکل برای پیدا کردن اطلاعات ساختاری، فعل و انفعالات بین نواحی کروموزومی را مورد بررسی قرار داده و آنها را استخراج می‌کند. داده‌های خام استخراج شده توسط این پروتکل به صورت داده‌های متنی شامل حروف A، C، T و G[1] است. منظور از حروف A، C، T و G همان بازهای

[1] adenine (A), cytosine (C), guanine (G), thymine (T)

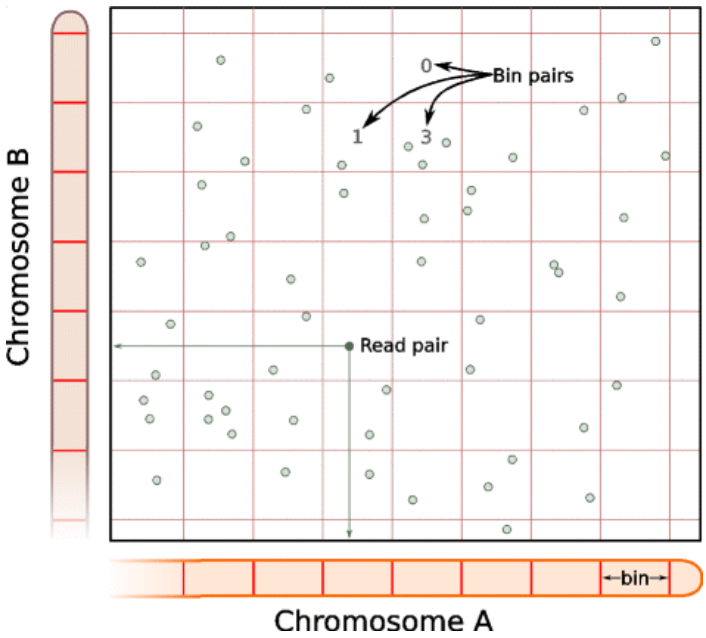
نوکلئوتیدی تشکیل دهنده یک رشته DNA است. داده خام استخراج شده توسط پروتکل Hi-C، توسط ابزارهای مختلفی پردازش شده و ماتریس مربوط به فعل و انفعالات بین نواحی کروموزومی تولید می‌شود [25] ,[24]. در واقع ورودی خام مورد استفاده توسط اکثر روش‌های پیش پردازش همان ماتریس مربوط به فعل و انفعالات است [28]–[26] ,[21]. این مساله را می‌توان در شکل ۱-۲ مشاهده نمود. در بخش (A) این شکل مرحله استخراج داده‌های Hi-C نمایش داده شده‌اند و در بخش (B) ماتریس تولیدی از این داده‌ها آورده شده است. در نهایت هم در بخش (C) این شکل پیش‌بینی ساختار کروموزومی از روی ماتریس فعل و انفعالات آورده شده است.



شکل ۱-۲ شکل مربوط به فرایند تولید داده‌های Hi-C از ابتدای فرایند استخراج فعل وانفعالات تا تولید ماتریسی از فعل و انفعالات [29]

ماتریس تشکیل شده از فعل و انفعالات، شامل مکان نواحی تشکیل دهنده فعل و انفعالات در یک کروموزوم و یا بین دو کروموزوم است. به عبارت دیگر برای هر فعل و انفعال مکان دو ناحیه تشکیل دهنده آن فعل و انفعال و همچنین برخی اطلاعات کنترلی ذخیره می‌شود. پس حذف برخی از فعل و انفعالات نامعتبر و نویز ساده، برای مشخص کردن فعل و انفعالات معنی‌دار و حذف نویز پیچیده نمی‌توان به صورت مستقیم از اطلاعات موجود دراین داده‌ها استفاده کرد. از این رو برای مشخص کردن خطا در این داده‌ها از افزایش اندازه ناحیه‌ها استفاده می‌شود. به عبارت دیگر با افزایش اندازه ناحیه‌ها می‌توان تاثیر خطا را بر فعل و انفعال بین دو ناحیه بزرگتر مشاهده کرد. لذا برای شناسایی نویز در داده‌های Hi-

C بعد نرمال‌سازی اولیه، اندازه ناحیه‌ها (بزرگتر و یا برابر با مقدار اولیه) مشخص می‌شوند. با مشخص شدن اندازه ناحیه‌ها با مفهومی به نام Read counts (Bin pairs) روبرو می‌شویم. در واقع Read counts تعداد فعل و انفعالات اولیه است که بین دو ناحیه بزرگتر وجود دارد. زمانی که اندازه ناحیه‌ها مشخص می‌شوند، مقدار Read counts را می‌توان محاسبه کرد و در نتیجه با استفاده از روش‌های مختلف، تاثیر نویز را بر مقدار Read counts هر فعل و انفعال مشخص کرد.



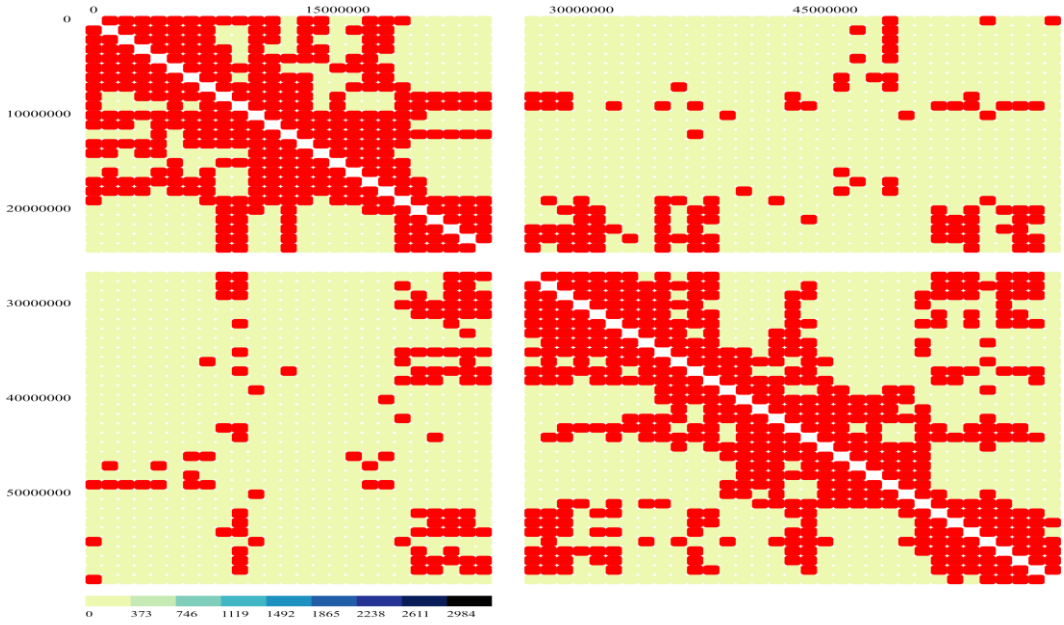
شکل ۲-۲ نمایی از مفاهیم استفاده شده در داده‌های Hi-C [30]

به عنوان مثال اگر یک کروموزوم شامل صد هزار جفت باز[1] باشد و این کروموزوم را به ناحیه‌هایی با اندازه پنج هزار جفت باز تقسیم می‌کنیم. در نتیجه می‌توان دریافت که این کروموزوم شامل ۲۰ ناحیه می‌باشد. حال اگر اندازه ناحیه‌ها را به بیست هزار جفت باز افزایش دهیم، این کروموزوم از ۵ ناحیه تشکیل خواهد شد که هر یک از این ناحیه‌ها شامل چهار ناحیه اولیه می‌باشند. در نتیجه فعل و انفعال تشکیل شده بین دو ناحیه بزرگتر برابر است با تعداد فعل و انفعالاتی که بین نواحی کوچکتر وجود دارند. به عبارت دیگرتعداد فعل و انفعالاتی که بین نواحی کوچکتر وجود دارد همان Read counts هر فعل و انفعال می‌باشد. مفاهیم ارائه شده را می‌توان در شکل ۲-۲ مشاهده نمود. در این شکل هر خانه

[1] Base pair

داخل ماتریس (شکل) نشان دهنده یک فعل و انفعال بین دو ناحیه کروموزومی می‌باشد. در داخل هر خانه ماتریس چندین نقطه وجود دارد که این نقاط نمایش دهنده فعل و انفعالاتی است که بین نواحی کوچکتر وجود دارند. به عبارت دیگر تعداد نقاط همان مقدار Read counts است. به طور کلی با توجه به مقدار Read counts می‌توان تاثیر نویز و خطا را بر داده‌های Hi-C مشخص نمود. در این پژوهش نیز از همین ایده برای شناسایی و حذف نویز استفاده کرده‌ایم.

علاوه بر این در این پژوهش با مفهوم دیگر به نام فعل و انفعالات معنی‌دار روبه‌رو هستیم. به طور کلی فعل و انفعالات معنی‌دار، فعل و انفعالاتی هستند که به صورت آماری معتبر (معنی‌دار) شناسایی شده‌اند. در واقع با استفاده از آزمون‌های آماری سعی می‌شود که درستی مقادیر تخمین زده شده هر Read counts را به صورت آماری بیان کرد. لذا تعداد، فعل و انفعالات معنی‌دار برای ارزیابی هر روشی که در این موضوع به صورت آماری حذف نویز انجام می‌دهد لازم است. همچنین لازم به ذکر است که علاوه‌بر ارزیابی درستی فعل و انفعالات، فعل و انفعالات معنی‌دار نشان دهنده فعل و انفعالات تاثیر گذار در ساختار کروموزومی است. به عنوان مثال با توجه به خانه‌های قرمز رنگ شکل ۳-۲ مشاهده می‌شود که فعل و انفعالات معنی‌دار شامل فعل و انفعالاتی با ناحیه‌های نزدیک به هم است. همچنین این فعل و انفعالات به توانسته‌اند تمام ناحیه‌های کروموزومی را پوشش دهند. لذا می‌توان از این فعل و انفعالات برای بررسی ساختار کروموزومی استفاده نمود.



شکل ۳-۲ نمایی از فعل و انفعالات معنی‌دار که به رنگ قرمز مشخص شده‌اند

۲-۲-۲ رگرسیون

تحلیل رگرسیون یک فرایند آماری برای تخمین روابط بین متغیرها است [32] ,[31]. این روش شامل تکنیک‌های زیادی برای مدل‌سازی و تحلیل متغیرها است. هدف تحلیل رگرسیون تعیین بهترین مدلی است که چگونگی ارتباط یک متغیر را با یک یا چند متغیر دیگر تعیین می‌کند. در این مساله متغیرها به دو دسته مستقل و وابسته تقسیم می‌شوند. متغیر وابسته عامل اصلی است که پژهشگران سعی در درک یا پیش بینی آن دارند. در هر مساله در مقابل متغیر وابسته، متغیرهای مستقل تعریف می‌شوند. طبق فرض هر مساله، متغیرهای مستقل عوامل اصلی تاثیر گذار بر متغیر وابسته هستند. تحلیل رگرسیون به پژوهشگران در فهم این موضوع که چگونه مقدار متغیر وابسته به تغییرات متغیرهای مستقل وابسته است، کمک می‌کند. به عبارت دیگر تحلیل رگرسیون تابعی از متغیرهای مستقل را برای محاسبه مقدار متغیر وابسته، تخمین می‌زند. برای مثال پیش‌بینی معدل ترم جاری یک دانشجو با توجه به نمرات چندین درس در ترم‌های قبل و یا سود حاصل از فروش محصولات یک شرکت با توجه به فروش و سود سال‌های گذشته از این نوع مسائل می‌باشند. تکنیک‌های زیادی برای انجام تحلیل

رگرسیون با توجه به پیچیدگی مساله توسعه داده شده است. روش‌های همچون رگرسیون خطی و رگرسیون پواسون.

در این پژوهش ابتدا از رگرسیون دوجمله‌ای منفی[1] استفاده کرده‌ایم. این روش در واقع تعیم یافته روش رگرسیون پواسون[2] هست که در آمار، رگرسیون پواسون یکی از انواع روش‌های تحلیل رگرسیون است. این روش زیر مجموعه‌ای از مدل‌های خطی تعمیم‌یافته است که برای تحلیل داده‌های حاصل از شمارش کاربرد دارد. از انجایی که این روش بر مبنای رگرسیون پواسون توسعه یافته است، در نتیجه این دو روش در بسیاری از منابع در کنار هم توضیح داده شده‌اند. تفاوت اساسی این روش با رگرسیون پواسن در فرض اولیه این روش‌ها است. در واقع در روش رگرسیون دوجمله‌ای منفی بر خلاف رگرسیون پواسون، فرض بسیار محدود کننده برابری واریانس با میانگین را ندارد. از این رو به دلیل نبود این فرض محدود کننده، روش رگرسیون دوجمله‌ای منفی بسیار پرکاربرد است.

۳-۲-۲ خوشه بندی

خوشه بندی در واقع به روش‌هایی گفته می‌شود که مجموعه ای از اشیاء را در گروه‌های مختلف به نحوی قرار می‌دهند. فرایند گروه‌بندی در این روش‌ها به شیوه‌ای انجام می‌شود که هر شی عضو یک گروه به اشیاء درون گروه خودش در مقایسه با دیگر گروه‌ها شباهت بیشتری داشته باشد. تا کنون الگوریتم‌های مختلف خوشه بندی توسعه یافته‌اند. یکی از روش‌های توسعه داده شده، روش K-Means است. روش K-Means یکی از ساده‌ترین الگوریتمِ‌های خوشه بندی است. با این وجود ممکن است در بسیاری از مسائل در دسته بهترین الگوریتم‌های خوشه بندی دسته‌بندی شود. الگوریتم K-Means یک الگوریتم تکرار شونده است که در واقع سعی می‌کند یک مجموعه n عضوی را به k زیر مجموعه افراز کند. در هر تکرار برای افراز یک مجموعه n عضوی این الگوریتم هر عضو را با میانگین هر خوشه مقایسه می‌کند که در نتیجه هر عضو مجموعه افراز می‌شود به خوشه‌ای که کمترین فاصله را دارا می‌باشد.

[1] negative binomial regression
[2] poisson regression

سپس بعد از هر بار تکرار میانگین هر خوشه دوباره محاسبه شده و دوباره فاصله هر عضو با میانگین خوشه‌ها مقایسه می‌شود. این الگوریتم تا جایی ادامه پیدا می‌کند که عضوهای هر زیر مجموعه تغییر نکنند. در این پژوهش از روش K-Means برای خوشه بندی داده‌ها بر اساس مکان کروموزومی آن‌ها استفاده کرده‌ایم. هدف ما در استفاده از خوشه بندی، پیدا کردن فعل و انفعالات همسایه بوده است. واقع هر خوشه به عنوان ورودی شبکه عصبی اتوانکدر استفاده شده است.

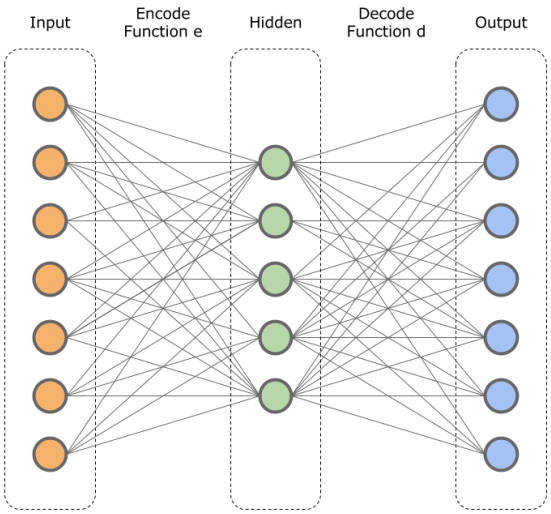
۲-۲-۴ شبکه عصبی

شبکه‌های عصبی مصنوعی، به عنوان یکی از پرکاربردترین و پویاترین زمینه‌های پژوهش در حوزه‌های مختلف محسوب می‌شوند. ایده اصلی شبکه‌های عصبی مصنوعی الهام گرفته از نحوه کار و یادگیری دستگاه عصبی زیستی است که در جانوران وجود دارد. در نتیجه مانند دستگاه عصبی، شبکه‌های عصبی از شمار زیادی واحدهای پردازشی به نام نورون تشکیل شده است. نورون‌ها در شبکه‌های عصبی به صورت لایه‌ای قرار گرفته‌اند. لازم به ذکر است که رفتار شبکه نیز وابسته به ارتباط بین اعضا است.

شبکه‌های عصبی مصنوعی در موضوعات مختلفی مانند مدل‌سازی، طبقه‌بندی و یادگیری عمیق کاربرد دارد که براساس نوع مساله ساختار شبکه عصبی تعیین می‌شود. همچنین لازم به ذکر است که آموزش در شبکه عصبی می‌تواند با نظارت و یا بدون نظارت باشد که این موضوع نیز بستگی به نوع مساله دارد.

همان‌طور که قبل به آن اشاره شد شبکه عصبی در موضوع یادگیری عمیق کاربرد بسیار زیادی دارد. به عبارت دیگر این دو موضوع به هم وابسته هستند. یادگیری عمیق زیر شاخه یادگیری ماشینی است که در آن بر مبنای مجموعه‌ای از الگوریتم‌ها سعی بر مدل کردن مفاهیم انتزاعی سطح بالا در دادگان را دارد. در واقع یادگیر عمیق بر اساس داده‌ها الگوهایی را تولید می‌کند. برای مدل کردن این داده‌ها، یادگیری عمیق از انواع مختلفی از شبکه‌های عصبی استفاده می‌کند [35]–[33]. یکی از این شبکه‌ها،

شبکه عصبی عمیق اتوانکدر[1] (شکل ۴-۲) است [36].

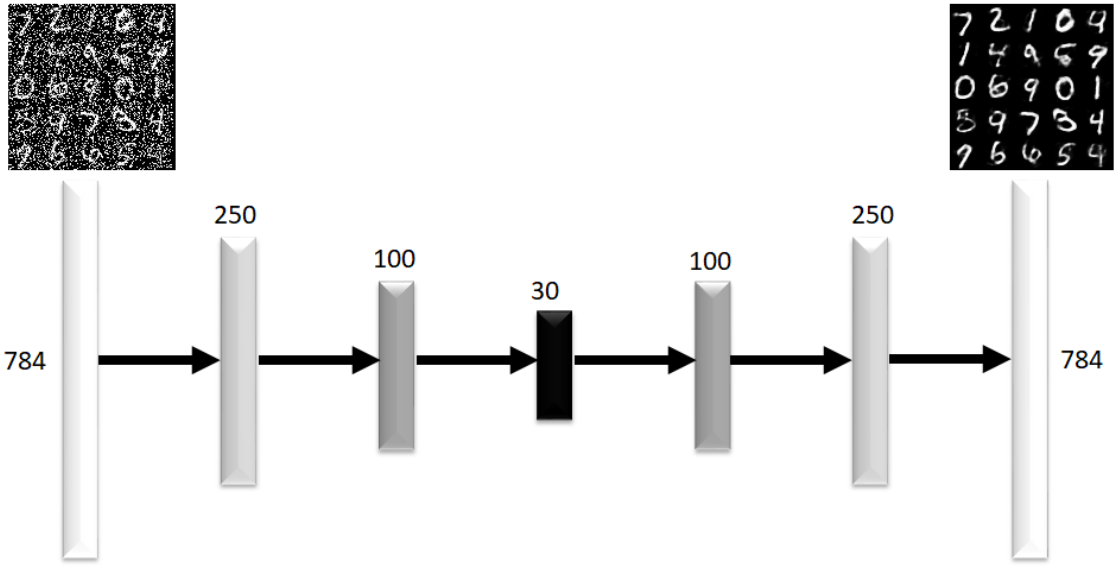


شکل ۴-۲ نمایی از شبکه عصبی عمیق اتوانکدر

شبکه عصبی عمیق اتوانکدر یک شبکه عصبی بدون نظارت است. این شبکه به صورت کلی دارای دو مرحله اساسی Encoding Input Data و Decoding Representations است. در مرحله اول این شبکه داده‌ی وروری را به مجموعه ای از ویژگی‌ها تبدیل می‌کند. یا به عبارت دیگر این شبکه داده‌های ورودی را به رمز تبدیل می‌کند. سپس در مرحله دوم، این شبکه سعی میکند که دوباره داده‌ی ورودی را براساس ویژگی‌های ایجاد شده در مرحله اول بازسازی نماید. در واقع مرحله اول ویژگی‌های اصلی داده‌ها را ایجاد می‌کند. در نتیجه با بازسازی این داده‌ها بر اساس این ویژگی‌ها سبب می‌شود که بخش زیادی از نویز موجود در داده‌ها حذف شوند. برای نمونه اگر یک تصویر با ۷۸۴ پیکسل داشته باشد (۲۸×۲۸) داشته باشیم و ساختار شبکه عصبی اتوانکدر را با پنج لایه به صورت شکل ۵-۲ تعریف کنیم. مشاهده می‌شود که در طول فرایند آموزش بخش Encoding (لایه اول شبکه به ۲۵۰ نورون به سمت لایه سوم شبکه با ۳۰ نورون) شبکه عصبی یاد می‌گیرد که چگونه ویژگی‌های مهم داده‌ها را استخراج می‌کند. سپس در مرحله آموزش بخش Decoding (لایه سوم شبکه به ۳۰ نورون به سمت لایه پنجم شبکه با

[1] Autoencoder

۲۵۰ نورون) شبکه عصبی یاد می‌گیرد که چگونه بر اساس این ویژگی‌ها دوباره داده را بازسازی نماید. با توجه به فرایند مدل کردن داده‌ها در این شبکه متوجه می‌شویم که روابط غیر خطی داده‌ها نیز مدل می‌شوند. به طور کلی از آنجایی که این شبکه عصبی قادر به یادگیری روابط غیر خطی می‌باشد در نتیجه این شبکه کاربرد بسیار زیادی در حذف نویز دارد. در این پژوهش به خاطر ماهیت بدون نظارت داده‌های Hi-C و نحوه مدل کردن داده‌ها در این شبکه، در این پژوهش از اتوانکدر این شبکه استفاده نموده‌ایم.



شکل ۵-۲ ساختار شبکه عصبی اتوانکدر با پنج لایه

۲-۳ مروری بر پیشینه تحقیق

تا کنون روش‌های مختلفی برای حذف نویز از داده‌های Hi-C ارائه شده است [39]–[37]. در این بخش چهار روش HiCNorm، Fit-Hi-C، GOTHiC و CHiCAGO به‌طور مختصر مرور خواهند شد.

۲-۳-۱ HiCNorm

یکی از اولین روش‌های توسعه یافته روش HiCNorm است [18]. این روش فرض می‌کند که دادگان دارای توزیع پواسن هستند. در آمار و احتمال توزیع پواسون یک توزیع احتمالاتی گسسته است که

احتمال اینکه یک حادثه به تعداد مشخصی در فاصله زمانی یا مکانی ثابتی رخ دهد را شرح می دهد. در واقع این روش، فرض می‌کند که احتمال رخ دادن فعل و انفعالات در فاصله مکانی ثابت (مکان ناحیه‌های کروموزومی) تابعی از توزیع پواسن است. لذا برای مدل کردن دادگان، روش HiCNorm از رگرسیون پواسون استفاده می‌کند. این روش با استفاده از رگرسیون پواسون و بر اساس مولفه‌های GC content و mappability تاثیر خطاهای سیستماتیک و نویز را مدل کرده و تخمینی از مقدار Read count فعل و انفعلات بین نواحی کروموزومی را محاسبه می‌کند. لازم به ذکر است که در اینجا مولفه mappability مقیاسی برای سنجیدن توانایی هم تراز کردن قعل و انفعالات به یک مکان منحصر به فرد در یک ژنوم است. همچنین مولفه GC content نشان دهنده نسبت تعداد مولکول‌های guanine و cytosine به تعداد کل مولکول‌های موجود در یک رشته DNA است.

با وجود اینکه روش HiCNorm یک مدل آماری را برای توصیف دادگان Hi-C تولید می‌کند. با این وجود این روش با مشکلات مختلفی روبرو است. به عنوان نمونه، این روش فقط مقدار نرمال شده Read count را محاسبه می‌کند. همچنین در این روش مقادیر Read count تخمین زده شده هر فعل و انفعال عموما از مقدار اولیه Read count کمتر است. لذا به طور دقیق نمی‌توان گفت مقادیر Read count به درستی تخمین زده شده‌اند. لازم به ذکر است که این روش بر خلاف بسیاری از روش‌های دیگر این روش برای شناسایی فعل و انفعالات معنی‌دار راهکاری در نظر گرفته نشده است.

۲-۳-۲ Fit-Hi-C

روش دیگری که برای بهبود روش HiCNorm توسعه یافته بود. روش Fit-Hi-C است [17]. این روش به صورت کلی دارای دو بخش اساس می‌باشد. در بخش اول، روش Fit-Hi-C از spline-fitting procedure استفاده می‌کند. الگوریتم spline fitting روشی است که در مسائل درون‌یابی کاربرد دارد. به عبارت دیگر این روش یکی از انواع روش‌هایی است که برای تعیین یک منحنی یا تابع ریاضی که بیشترین شباهت را به داده‌ها داشته باشد، استفاده می‌شود. در روش Fit-Hi-C برای بدست آوردن تابعی

که بتواند داده‌ها را به خوبی مدل کند و شرح دهد، از روش spline fitting استفاده شده است. در ادامه این فرایند ابتدا برخی از فعل و انفعالات توسط مدل بدست آمده حذف می‌شوند. سپس توسط برایب بهبود مدل بدست آمده توسط spline fitting، این روش فرض می‌کند که دادگان از توزیع دوجمله‌ای منفی پیروی می‌کنند. در نتیجه برای مدل کردن و نرمال سازی داده‌ها این روش از رگرسیون دو جمله‌ای منفی استفاده می‌کند. لازم به ذکر است که این روش دارای مشکلات مختلفی است، از جمله اینکه این روش برای اطلاعات مربوط به فعل و انفعالات بین کروموزومی[1] قابل استفاده نیست. همچنین این روش نمی‌تواند از انواع پروتکل‌های Hi-C پشتیبانی کند.

۲-۳-۳ GOTHiC

یکی دیگر از روش‌های توسع داده شده، روش GOTHiC است [19]. این روش ابتدا فرض می‌کند نویز و خطاهای سیستماتیک بر هر دو ناحیه یک فعل و انفعال تاثیر گذاشته است. سپس احتمال رخ دادن هر فعل و انفعال و در پی آن تخمینی از مقدار Read counts را بر اساس فرض ذکر شده برای هر فعل و انفعال محاسبه می‌کند. این روش با استفاده از آزمون دوجمله‌ای[2] و بر اساس مقدار تخمینی و مقدار اولیه Read counts، فعل و انفعالات معنی‌دار را پیدا می‌کند. در واقع این روش برای شناسایی اینکه آیا فعل و انفعال دو ناحیه از کروموزوم معنی‌دار است یا نه از آزمون دوجمله‌ای پیرو روش Benjamini–Hochberg multiple testing correction استفاده می‌کند.

در روش ارائه شده GOTHiC مشکلاتی متعددی وجود دارد. اصلی‌ترین مشکل این روش استفاده نکردن از برخی ویژگی‌ها مانند معیار فاصله، GC content و mappability می‌باشد. همچنین می‌توان به متکی بودن این روش تنها به پارامترهای محدودی مانند مقدار Read counts در شناسایی خطاهای سیستماتیک و پشتیبانی نکردن این روش از انواع داده‌های Hi-C به عنوان مشکلات اساسی اشاره کرد. لازم به ذکر است این روش مزایای زیادی نسبت به دو روش قبل دارد. یکی از مزایای این روش نسبت

[1] inter interactions
[2] Binomial test

به دو روش قبل سرعت پردازش دادگان است. در واقع برای دادگان یکسان این روش با سرعت بیشتری فعل وانفعالات معنی‌دار را پیدا می‌کند. همچنین لازم به ذکر است که این روش برخلاف روش‌های Fit-Hi-C و HiCNorm از اطلاعات مربوط به دو ناحیه کروموزومی با کروموزوم‌های غیر یکسان پشتیبانی می‌کند.

۲-۳-۴ CHiCAGO

یکی دیگر از روش‌های توسعه یافته، روش CHiCAGO است [22]. این روش یک فرایند آماری طراحی شده برای شناسایی فعل و انفعالات معنی‌دار برای داده‌های Capture Hi-C می‌باشد. در این روش فرض بر این است که فعل وانفعالات معنی‌دار بر اساس وابستگی فاصله (تابع فاصله) قابل شناسایی هستند. بنابراین در CHiCAGO روشی مبتنی بر فاصله برای شناسایی فعل وانفعالات معنی‌دار ارائه شده است. روش ارائه شده در CHiCAGO را می‌توان به سه بخش مهم محاسباتی تقسیم کرد. لازم به ذکر است که این سه بخش قسمتی از روش ارائه شده می‌باشند:

- مدل‌سازی

در این بخش به مدل کردن تعداد فعل و انفعالات مشاهده شده پرداخته می‌شود. برای این منظور در این روش ابتدا read counts ناشی از حرکت براونی را با استفاده رگرسیون دو جمله‌ای منفی مدل کرده که در آن سطوح تخمینی، تابعی از فاصله ژنومی است. به غیر از مدل اول در این روش نویز فنی ناشی از خطاهای آزمایشگاهی با استفاده از توزیع Poisson (فرض می‌شود که این خطاها دارای توزیع Poisson هستند که با استفاده از trans-chromosomal counts تخمین زده می‌شود) مدل می‌شوند. این دو قسمت به صورت مجزا مدل شده و حاصل این دو قسمت به صورت یک توزیع Delaporte ترکیب می‌شوند. بعد از بدست آمدن توزیع Delaporte در این روش مقادیر p-values هر یک از جفت ناحیه‌ها یا به عبارت دیگر هر فعل و انفعال محاسبه می‌شوند.

- اصلاح مقادیر p-values

با محاسبه مقادیر p-values توسط این روش، می‌توان با اعمال یک حد آستانه بر روی p-values ، فعل وانفعالات معنی‌دار را شناسایی کرد. با این وجود بخش بزرگی از فعل و انفعالات معنی‌دار شناسایی شده، long-range interactions می‌باشند. به عبارت دیگر بخش بزرگی از فعل و انفعالات معنی‌دار شناسایی شده دارای فاصله کروموزومی زیاد هستند. بنابر این روش طبق فرض اصلی مساله که در فصل اول بیان شد. مقادیر p-value را اصلاح می‌کند. به عبارت دیگر در روش CHiCAGO از استراتژی p-value Weighting استفاده شده است. این استراتژی وزن فعل و انفعالاتی که فاصله کمی دارند و وزن فعل و انفعالاتی که فاصله زیادی دارند را به ترتیب زیاد و کم می‌کند. بعد از محاسبه وزن‌ها با استفاده از این وزن‌ها مقادیر p-values اصلاح می‌شوند. به طور کلی می‌توان گفت که این روش به فرض اصلی مساله را مورد توجه قرار داده است.

- محاسبه امتیاز

بعد از اصلاح مقادیر p-values، در این روش معیاری با نام امتیاز (Score) تعریف شده است که مقدار آن طبق فرمول (۲-۱) محاسبه **می‌شود.**

$$Score_{ij} = \max\left(0, -logQ_{ij} - logW_{max}\right) \qquad (۱-۲)$$

در معادله بالا w_{max} برابر با حداکثر weight قابل دسترسی است و Q_{ij} نیز به صورت زیر تعریف می‌شود:

$$Q_{ij} = \frac{P_{ij}}{W_{ij}} \qquad (۲-۲)$$

در روش CHiCAGO با استفاده از امتیاز تعریف شده، به تعیین آستانه (در مقاله CHiCAGO مقدار این آستانه برابر است با $Score \geq 5$) و شناسایی فعل وانفعالات معنی‌دار پرداخته می‌شود. به عبارت دیگر هر فعل وانفعالی که امتیازی بیشتر یا برابر با حد آستانه داشته باشد به عنوان یک فعل وانفعال معنی‌دار شناخته می‌شود. این روش نیز دارای مشکلات مختلفی است که می‌توان به مشکل کند بودن

این روش و همچنین پشتیبانی این روش از یک نوع خاص از داده‌های Hi-C با پیچیدگی کمتر است، اشاره کرد.

۲-۴ جمع‌بندی

تا کنون روش‌ها مختلفی برای نرمال‌سازی داده‌های Hi-C و همچنین شناسایی داده‌های معنی‌دار توسعه داده شده‌اند. هر یک از روش‌های بررسی شده در این بخش دارای معایب خاص خود هست. با این وجود برخی مشکلات بین این روش‌ها مشترک هستند. یکی از این مشکلاتی که در عموم روش‌های ارائه شده وجود دارد، این است که این روش‌ها از برخی مولفه‌های مهم در مدل‌سازی دادگان استفاده نکرده‌اند. در واقع عموما در روش‌های ارائه شده از روش‌های آماری برای مدل کردن دادگان با توجه به مولفه‌هایی مانند فاصله، استفاده می‌کنند. در نتیجه مولفه‌های محلی تاثیر گذار بر فعل و انفعالات مورد بررسی قرار نمی‌گیرند. یکی دیگر از مشکلات مطرح شده، عدم پشتیبانی این روش‌ها از انواع مختلف داده‌های Hi-C است. به طور کلی از آنجایی که مشکلات مختلفی در روش‌های ارائه شده وجود دارد. لذا در این پژوهش سعی شده است که به حل این مشکلات پرداخته می‌شود.

فصل ۳ روش پیشنهادی

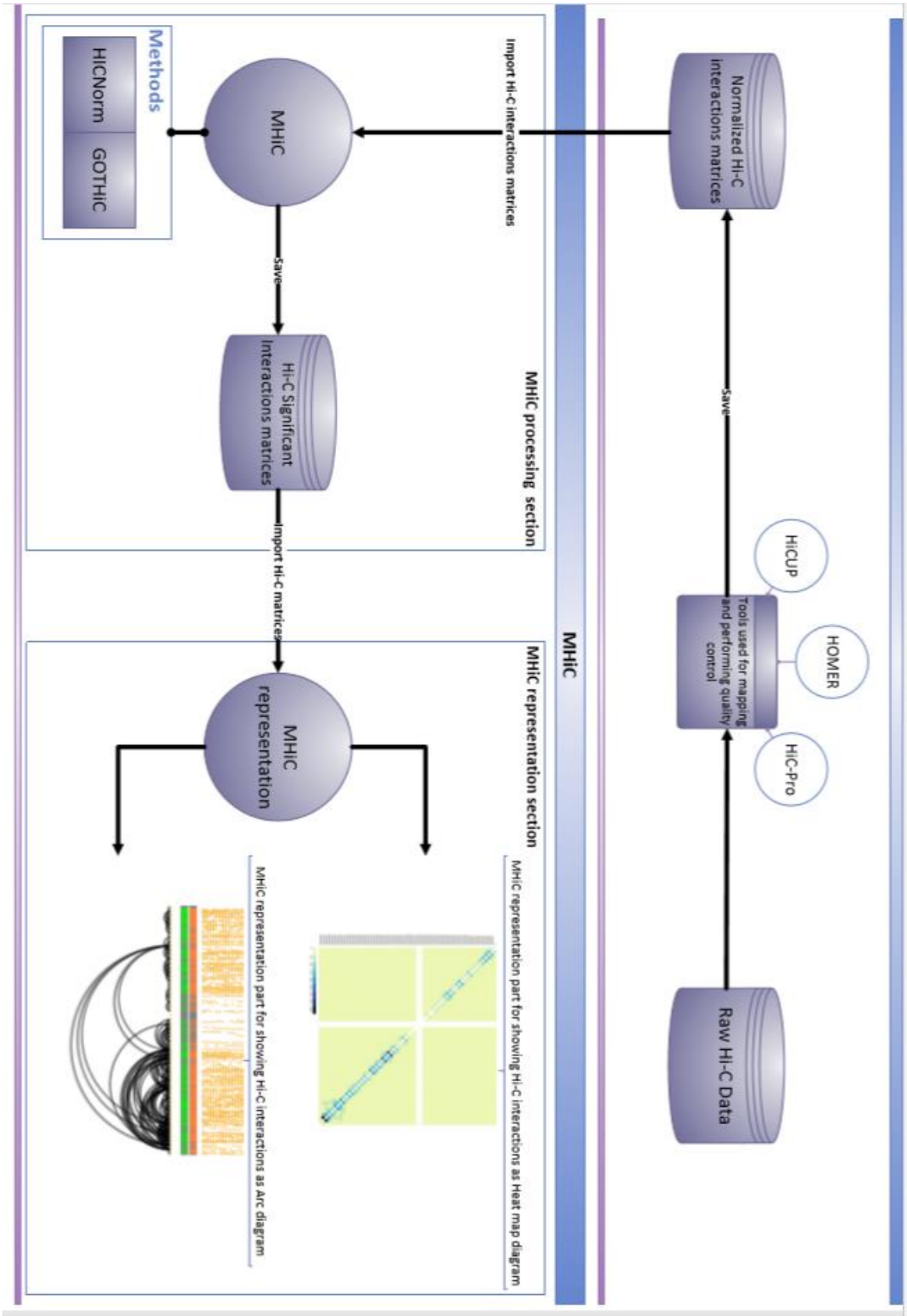
۳-۱ پیش‌درآمدی بر روش پیشنهادی

پروتکل Hi-C همان طور که در فصل‌های قبل بیان شده اطلاعاتی را در مورد ساختار فضایی کروموزوم‌ها به ما می‌دهد. در این پروتکل نویز مشکل اساسی است. در نتیجه هدف اصلی ما در این پژوهش حذف نویز و خطاهای سیستماتیک از داده‌های Hi-C بوده است. با این حال توسعه یک ابزار نرمال سازی برای استفاده دیگر پژوهشگران نیز قسمتی از اهداف ما در این پژوهش بوده است. از آنجای که هدف ما در این پژوهش دو مرحله‌ای بوده است. در این فصل ابتدا ابزار توسعه داده شده را همراه با روش‌های مورد استفاده در این ابزار شرح داده می‌شود. سپس در ادامه بخش اول، روشی را که برای نرمال سازی داده‌های Hi-C توسعه داده‌ایم را به صورت مرحله به مرحله توضیح می‌دهیم.

۳-۲ ابزار MHiC

یکی از اهداف ما در این پژوهش تولید ابزاری برای حذف نویز و شناسایی فعل و انفعالات معنی‌دار بوده است. در نتیجه ما ابزاری با نام MHiC را برای شناسایی نویز و خطاهای سیستماتیک در محیط R توسعه داده‌ایم. هدف ما از توسعه این ابزار، ارائه ابزاری جامع برای حذف نویز و همچنین بصری‌سازی داده‌های Hi-C است. به طور کلی سعی کرده‌ایم که روش‌های موجود برای حذف نویز در این داده‌ها را یکجا و در یک ابزار پیاده‌سازی نماییم. این ابزار به طور کلی با توجه به شکل ۳-۱ دارای سه بخش اصلی است:

۱. دریافت و هماهنگ سازی داده ورودی
۲. شناسایی و حذف نویز از دادگان
۳. بصری سازی داده‌های Hi-C



شکل ۱-۳ نمایی از مراحل و عملکرد ابزار توسعه یافته

از آنجایی که سیستم مطرح شده باید توانایی پردازش داده‌های حاصل از پروتکل‌های مختلف Hi-C را داشته باشد، و همچنین به دلیل ساختارهای متفاوتی که در داده‌ها وجود دارد، در بخش اول و ابتدایی این ابزار، ماژول‌هایی را برای پردازش داده‌های ورودی توسعه داده‌ایم. این ماژول‌ها ساختارهای مختلف ورودی را به ساختار استاندارد و نرمال شده سیستم تبدیل می‌کنند. به عبارت دیگر داده‌های حاصل از پروتکل‌های مختلف Hi-C برای بخش شناسایی نویز و خطاهای سیستماتیک آماده می‌شوند. در ادامه فرایند حذف نویز توسط این ابزار، داده‌ها در بخش شناسایی نویز و خطاهای سیستماتیک با استفاده از روش‌های GOTHiC، HiCNorm و FitHiC که در ابزار پیاده‌سازی شده‌اند، ارزیابی شده و خطای موجود در داده‌ها حذف می‌شوند. در پایان این سیستم، خروجی بخش دوم یا به عبارت دیگر بخش شناسایی نویز و خطاهای سیستماتیک توسط بخش بصری‌سازی، نمایش داده می‌شود. از نظر مفهومی این ابزار دارای دو بخش اساسی محاسبات و نمایش خروجی می‌باشد که در ادامه شرح داده می‌شوند.

۳-۲-۱ بخش محاسباتی

منظور از بخش محاسباتی ابزار همان بخش اصلی ابزار است که با استفاده از محیط R توسعه داده شده است. در این ابزار علاوه بر روش پیشنهادی، روش‌های GOTHiC، HiCNorm و FitHiC را نیز برای حذف نویز و شناسایی فعل و انفعالات معنی‌دار پیاده‌سازی نموده‌ایم. در نتیجه کاربر برای پردازش داده‌ها می‌تواند از یکی از این روش‌های پیاده‌سازی شده استفاده کند. لازم به ذکر است که هر یک از روش‌های پیاده‌سازی شده در این ابزار بر خلاف پیاده‌سازی‌های موجود، توانایی دریافت ورودی از منابع HiC-Pro، HOMER، HiCUP و همچنین ورودی طراحی شده برای روش HiCNorm را دارا می‌باشند. در نتیجه این ابزار بازه گسترده‌تری از داده‌های Hi-C نسبت به پیاده‌سازی اولیه روش‌های پوشش می‌دهد. در واقع در این ابزار سعی کرده‌ایم که ساختارهای مختلف دادگان Hi-C را با روش‌های موجود هماهنگ شوند.

۱-۱-۲-۳ ورودی

همان طور که در بخش قبل به آن شاره شد. این ابزار توانایی دریافت ورودی از منابع HiC-Pro، HOMER، HiCUP و همچنین ورودی طراحی شده برای روش HiCNorm را دارا می‌باشد. هر یک از این منابع در واقع خود ابزاری برای پردازش و سازی داده‌های خام Hi-C می‌باشند. داده‌های خام Hi-C در واقع داده‌های متنی هستند که رشته‌های کروموزم در آن با حروف A، T، C و G همراه با کنترل‌های مربوطه ذخیره شده‌اند. در نتیجه این ابزارها عمل پیش‌پردازش را برای ابزار ما انجام می‌دهند. مشکلی که در این ابزارها وجود دارد یکسان نبودن خروجی این ابزارها می‌باشد به عبارت دیگر هر یک از این ابزارها ساختار خروجی منحصر به فرد خود را دارا می‌باشد که در ادامه خروجی این ابزارها (ورودی ابزار ما) به صورت خلاصه شرح داده شده است.

۱. HiC-Pro

ابزار HiC-Pro یک ابزار کامل است که علاوه بر پروتکل اصلی Hi-C، از پروتکل‌های مبتنی بر پروتکال Hi-C پشتیبانی می‌کند [27]. این ابزار برای پردازش داده‌های خام Hi-C و نرمال‌سازی اولیه توسعه یافته است. در واقع این ابزار دادگان خام را به عنوان ورودی گرفته و خطاهایی مانند self ligations را از دادگان حذف می‌نماید. خروجی این ابزار به صورت ماتریسی از فعل و انفعالات نرمال شده است. لازم به ذکر است که این ابزار توانایی تولید خروجی در ساختارهای مختلف را دارد که در این ابزار از ساختاری با دو ماتریس خروجی استفاده کرده‌ایم. یکی از ماتریس‌ها اطلاعات مربوط به هر فعل و انفعال، در آن قرار دارد. در ماتریس دیگر اطلاعات مربوط به نواحی موجود (این ماتریس محل قرارگیری هر ناحیه را در کروموزوم نشان می‌دهد.) در داده‌های Hi-C موجود است. لازم به ذکر است که این روش به دلیل استفاده از پردازش موازی سرعت بیشتری نسبت به ابزارهای مشابه دارد.

۲. HiCUP

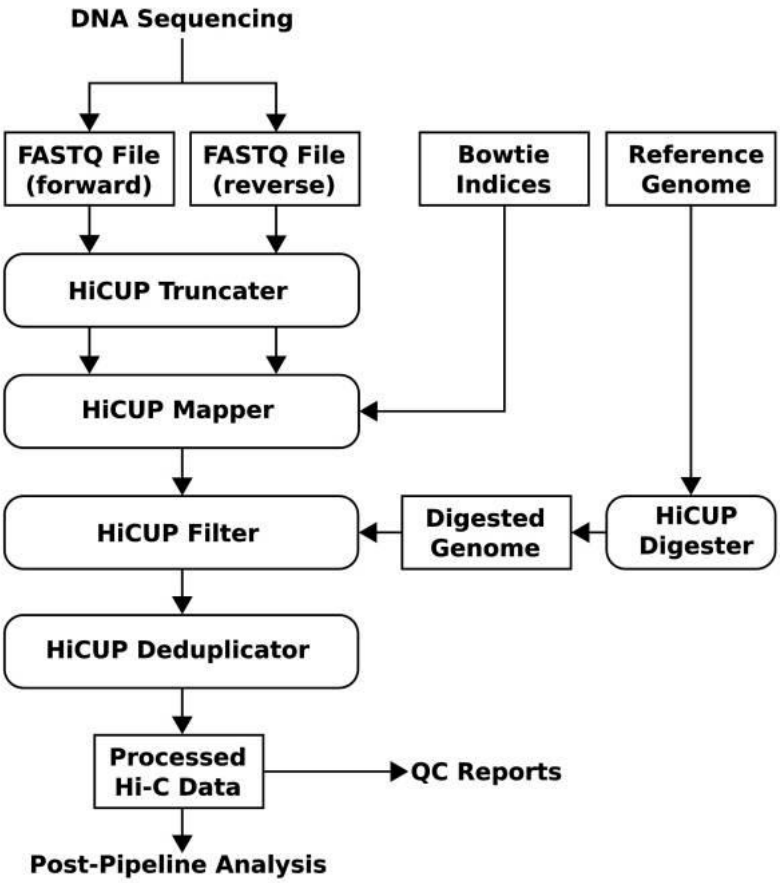
ابزار HiCUP یک ابزار کامل مانند ابزار HiC-Pro برای پردازش داده‌های خام Hi-C و نرمال سازی

اولیه داده می‌باشد [21]. با توجه به شکل ۲-۳ مشاهده می‌شود که این ابزار مانند ابزار HiC-Pro دادگان خام را به عنوان ورودی گرفته و خطاهایی مانند self ligations را از دادگان حذف کرده و در نهایت خروجی نرمال شده را برای آنالیز دادگان Hi-C آماده می‌کند. در واقع خروجی این ابزار نیز ماتریسی از فعل و انفعالات نرمال شده است. خروجی این ابزار نیز مانند ابزار HiC-Pro از دو ماتریس برای توضیح فعل و انفعالات و مکان ناحی‌ها در کروموزوم تشکیل شده است. با این وجود خروجی این ابزار از نظر ساختاری با خروجی HiC-Pro تفاوت دارد. لازم به ذکر است که این ابزار نسبت به ابزار HiC-Pro سرعت کمتری در انجام فرایند دارد.

۳. MHiC و HiCNorm

در روش HiCNorm برای پردازش داده‌ها از ورودی با ساختار خاصی استفاده می‌شود. ورودی طراحی شده برای روش HiCNorm دارای دو ماتریس برای اطلاعات فعل و انفعالات و اطلاعات تکمیل کننده می‌باشد. ماتریس اول شامل اطلاعات مربوط به هر فعل و انفعال است و ماتریس دوم شامل اطلاعات محل قرار گیری ناحیه‌ها و همچنین برخی اطلاعات تکمیلی (fragment length، GC content و mappbility) می‌باشد.

زمانی که ساختار دادگان شبیه به هیچ یک از ساختارها توصیف شده نباشد. از ورودی MHiC استفاده می‌شود. در واقع ورودی MHiC، ورودی طراحی شده ما برای زمانی است که کاربر داده‌ای با ساختار متفاوت دارد که دراین صورت کاربر با تغییر ساختار داده خود به ساختار ورودی MHiC می‌تواند داده خود را به ابزار MHiC به عنوان ورودی وارد کند.



شکل ۲-۳ فلوچارت مربوط به ابزار HiCUP [21]

۲-۲-۳ روش‌ها و الگوریتم‌های مورد استفاده

در این بخش ما به شرح روش‌های مورد استفاده برای حذف نویز در این ابزار می‌پردازیم. لازم به ذکر است این که روش‌ها در فصل پیشینه تحقیق معرفی شده‌اند.

۱. HiCNorm

این روش برای محاسبه مقدار نرمال Read Count هر فعل و انفعال ابتدا فرض می‌کند که این مقادیر دارای توزیع پواسن هستند [18]. سپس در این روش مقدار تخمینی Read Count برای هر فعل و انفعال به صورت زیر محاسبه می‌شود (شکل ۴-۳).

$$\mathbf{e^i_{jk}} = \frac{\mathbf{u^i_{jk}}}{\mathbf{t^i_{jk}}} \qquad (۱-۳)$$

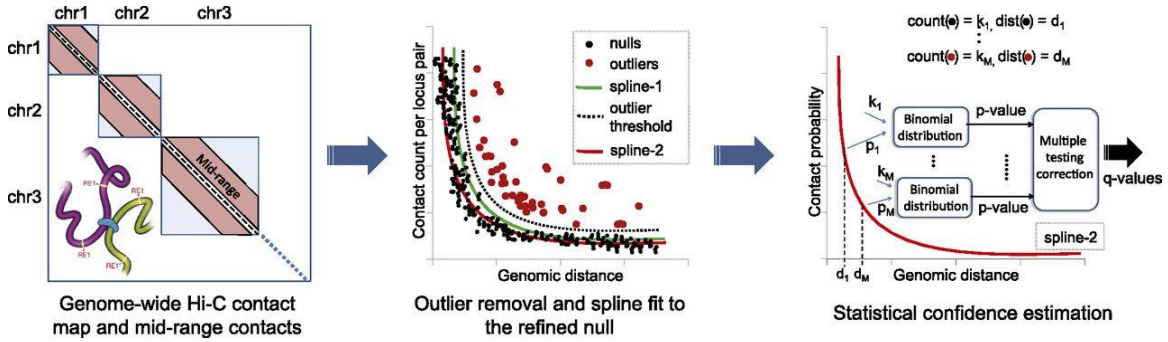
در این فرمول u و e به ترتیب نشـان دهنده مقدار Read Count اولیه و مقدار تخمینی این روش برای هر فعل و انفعال بین ناحیه‌های j و k در کروموزوم i می‌باشـــند. همچنین مقدار t (میانگین توزیع پواسن برای هر فعل و انفعال) به صورت زیر محاسبه می‌شود.

$$\mathbf{t_{jk}^{i} = exp[\beta_{0}^{i} + \beta_{len}^{i} lg(x_{j}^{i} x_{k}^{i}) + \beta_{gc}^{i} lg(y_{j}^{i} y_{k}^{i}) + lg(z_{j}^{i} z_{k}^{i})]} \qquad (۲-۳)$$

در این فرمول x، y و z به ترتیب نشـــان دهنده مقادیر مولفه‌های effective length، GC content و mappability هر فعل و انفعال می‌باشـــند. همچنین در این فرمول ضـــرایب β_{len}^{i} و β_{gc}^{i} نشـــان دهنده تاثیر نویز بر هر مولفه می‌باشد که بر اساس مقادیر مولفه‌های effective length، GC content تخمین زده می‌شوند.

۲. Fit-Hi-C

روش Fit-Hi-C از spline-fitting procedure برای پیدا کردن مدلی از داده‌ها استفاده می‌کند (شکل ۳-۳) [17]. در واقع این روش با استفاده از روش spline fitting دادگان را مدل کرده و حد آستانه‌ای برای آنها تعیین می‌کند. در نتیجه با اســـتفاده از از این حد آســـتلنه برخی از فعل و انفعالات حذف می‌شـــوند. در ادامه فرایند حذف نویز این روش دادگان نرمال شـــده توســـط روش spline fitting به صـــورت توزیع دو جمله‌ای منفی توزیع شـــده‌اند. لذا این روش این دادگان را با اســـتفاده رگرســـیون دوجمله‌ای منفی مدل می‌کند و در ادامه مقادیر p-value و در پی آن q-value هر یک از جفت ناحیه‌ها توسط این روش محاسبه می‌شوند. عمل شناسایی فعل و انفعالات معنی‌دار و حذف نویز نیز با استفاده از این مقادیر انجام می‌شود.



شکل ۳-۳ نمایی از نحوه انجام فرایند حذف نویز در روش Fit-Hi-C [17]

۳. GOTHiC

همان طور که در فصل پیشینه تحقیق به ان اشاره شد، روش GOTHiC برای شناسایی اینکه آیا فعل و انفعال دو ناحیه از کروموزوم معنی‌دار است یا نه از آزمون دوجمله‌ای پیرو روش Benjamini–Hochberg multiple testing correction استفاده می‌کند (شکل ۳-۴) [19]. این روش ابتدا فرض می‌کند که نویز و خطاهای سیستماتیک بر هر دو ناحیه یک فعل و انفعال تاثیر گذاشته است. در نتیجه در این روش احتمال مشاهده مقدار Read Count را برای هر فعل و انفعال به واسطه random ligations محاسبه می‌شود. بر اساس این توضیحات در این روش ابتدا مقدار relative coverage برای هر ناحیه به صورت زیر محاسبه می‌شود. این مقدار در واقع همان Coverage error برای هر فعل و انفعال است. به طور کلی این مقدار نشان دهنده این موضوع است که چه میزان جمعیت هدف (real read counts) با جمعیت مورد نظر نمونه برداری (raw read counts) همخوانی نداشته است. همچنین در این فرمول reads به مجموع مقدار Read count همه فعل و انفعالاتی که یک ناحیه در آن‌ها شرکت دارد، اشاره دارد.

$$\boldsymbol{relativecoverage_j = \frac{readsj}{N}} \qquad (۳-۳)$$

سپس احتمال اینکه هر فعل و انفعال نتیجه‌ای نادرست از اتصال دو ناحیه باشد محاسبه می‌شود.

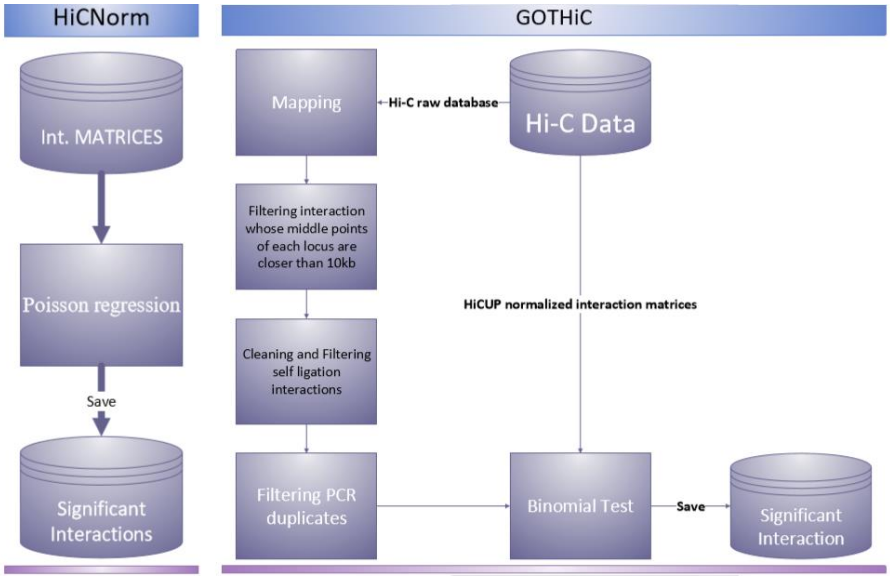
$$\boldsymbol{p_{jh} = relativecoverage_j \times relativecoverage_h} \qquad (۴-۳)$$

در آخر هم احتمال اینکه مقدار Read Count بین دو ناحیه مشاهده شده برابر یا بیشتر از مقدار read count تخمینی (n) باشد با استفاده از binomial cumulative density به صورت زیر بدست

می‌آید. لازم بذکر است که مقدار محاسبه شده در این فرمول برای سنجش درستی مقدار تخمینی read counts مورد استفاده قرار می‌گیرد. لازم به ذکر است که در این فرمول N به تعداد ناحیه‌های کروموزومی اشاره دارد.

$$pval_{jh} = 1 - \sum_{i=0}^{n_{jh}-1} \binom{N}{i} (p_{jh})^i (1-p_{jh})^{N-i} \qquad (۵-۳)$$

به طور کلی می‌توان گفت GOTHiC با استفاده از آزمون دوجمله‌ای، دو ناحیه از کروموزومی را که به طور قابل توجهی تعداد فعل وانفعالات بیشتری نسبت به شانس در آزمایش‌های Hi-C دارد را شناسایی می نماید و آن را به عنوان فعل و انفعال معنی‌دار معرفی می‌کند.



شکل ۴-۳ فلوچارت مربوط به روش‌های GOTHiC **و** HiCNorm

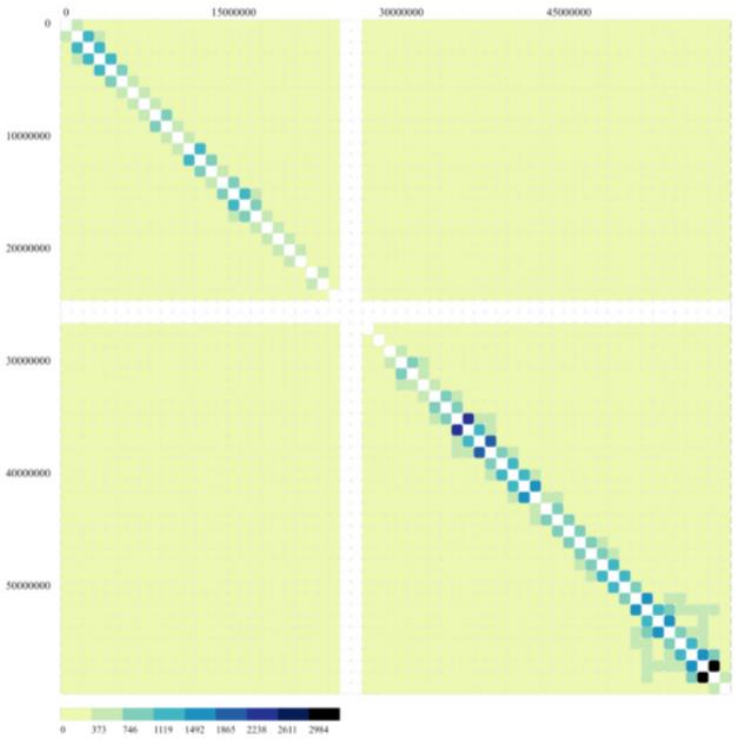
۳-۲-۳ خروجی

خروجی ابزار MHiC داده Hi-C نرمال سازی شده می‌باشد. این خروجی به صورت یک فایل متنی با فرمت csv ذخیره می‌شود. این فایل شامل اطلاعات هر فعل و انفعال مانند اطلاعات ناحیه‌های شرکت کننده در فعل و انفعال، مقدار اولیه read counts و مقدار تخمینی read counts است. در ادامه بعد از ذخیره خروجی ابزار MHiC، از خروجی می‌توان در بخش دیگر ابزار (بخش نمایش داده‌های Hi-C) استفاده نمود و خروجی را با استفاده از نمودارهای مختلف تعبیه در ابزار شده نمایش داد.

۳-۲-۴ بخش نمایش داده‌های Hi-C

در قسمت اول این بخش برای ابزار MHiC یک رابط کاربری برای استفاده راحت‌تر از بخش محاسباتی توسعه داده‌ایم. این بخش به صورت کد JavaScript توسعه داده شده است و سپس به محیط R متصل شده است.

در قسمت دیگر این بخش برای درک بهتر کاربر از داده‌های خروجی بخش محاسباتی ابزار، خروجی را با استفاده از Contact map Diagram (شکل ۳-۵) و Arc Diagram (شکل ۳-۶) نمایش می‌دهیم. این بخش نیز با استفاده JavaScript توسعه داده شده است و توانایی نمایش فعل و انفعالات معنی‌دار را دارا می‌باشد. در نتیجه کاربر می‌تواند به جای اینکه خروجی را به یک ابزار دیگر برای نمایش منتقل کند از این بخش برای نمایش خروجی استفاده کند.

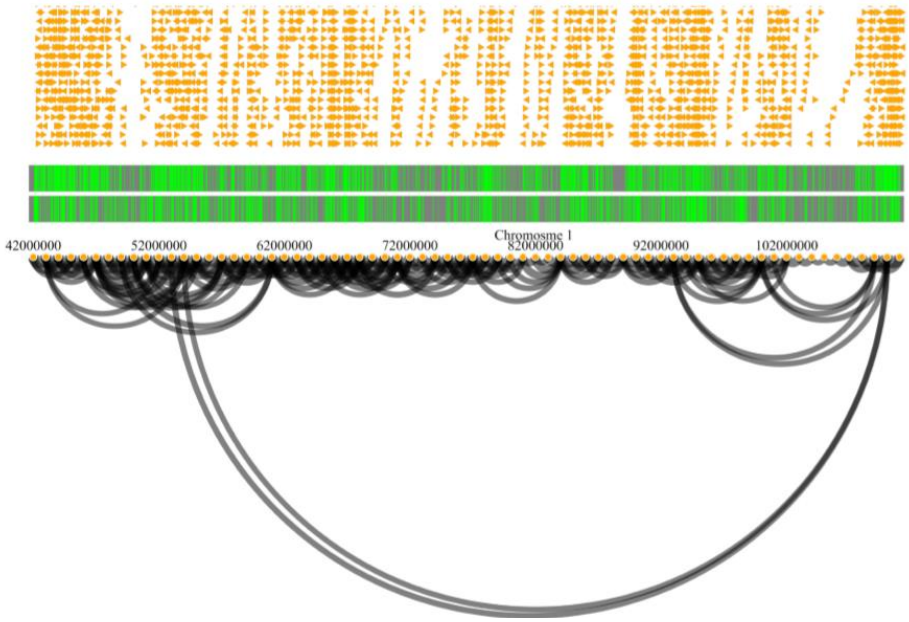


شکل ۳-۵ نمودار Contact map Diagram تولید شده توسط ابزار MHiC

۳-۳ مدل ترکیبی برای حذف نویز

همان‌طور که در فصل قبل به آن اشاره شد تا کنون روش‌های آماری زیادی برای نرمال‌سازی داده‌های

Hi-C توسعه یافته‌اند. این روش‌ها ابتدا سعی بر مدل‌سازی داده‌ها و بدست آوردن مقدار Read Count بر اساس مدل ساخته شده دارند. سپس بعد از مدل کردن و نرمال سازی داده‌ها بر اساس مقادیر اولیه و تخمینی Read Count، فعل و انفعالات معنی‌دار را شناسایی می‌کنند. روشی پیشنهادی نیز از همین مراحل پیروی می‌کند با این تفاوت که روش پیشنهادی در بخشی از فرایند مدل‌سازی از شبکه عصبی اتوانکدر استفاده می‌کند. روش پیشنهادی ابتدا دادگان را با استفاده از یک روش آماری مدل می‌کند. سپس این مدل آماری بدست آمده از دادگان را با استفاده از شبکه عصبی اتوانکدر بر اساس ویژگی‌های بهبود می‌بخشد. در ادامه برای هر فعل و انفعال بر اساس مدل بدست آمده از شبکه عصبی مقدار تخمینی Read Counts محاسبه می‌شود. در این روش همانند روش‌های گذشته بر اساس مقادیر اولیه و تخمینی Read Counts ، فعل و انفعالات معنی‌دار مشخص می‌شوند. لازم به ذکر است که علاوه بر ابزار توسعه داده شده، روش پیشنهادی نیز با نام MHiC توسعه داده‌ایم.



شکل ۶-۳ نمودار Arc Diagram **تولید شده توسط ابزار** MHiC

۳-۳-۱ طرح کلی

برای حذف نویز از داده‌های Hi-C در روش پیشنهادی، دو رویکرد محلی و سراسری را با هم ادغام کرده‌ایم. رویکرد سراسری به معنای آن است که برای نرمال‌سازی دادگان از برخی مولفه‌ها مانند مولفه

فاصله استفاده می‌شود. در این رویکرد تاثیر برخی مولفه‌ها بر روی کل دادگان مورد بررسی قرار می‌گیرد. به عبارت دیگر رویکرد سراسری به مولفه‌هایی اشاره دارد که بر تمام فعل و انفعالات موجود تاثیر گذارند. در این پژوهش رویکرد سراسری شامل محاسبه مقدار تخمینی read count بر اساس تاثیر مولفه فاصله بر هر فعل و انفعال است. برای محاسبه مقدار تخمینی read count در این پژوهش از روش ارائه شده در GOTHiC استفاده کرده‌ایم. در واقع در روش پیشنهادی با استفاده از بخشی از الگوریتم GOTHiC دادگان را مدل کرده و مقدار read count را تخمین می‌زنیم. مقادیر محاسبه شده بر اساس مدل آماری در ادامه فرایند حذف نویز، نقش مهمی در آموزش شبکه دارند. در واقع شبکه اتوانکدر سعی دارد مقدار اولیه را به مقدار تخمین زده شده نزدیک کند.

رویکرد محلی در این پژوهش به معنای مشاهده و بررسی نویز در گروهی از فعل و انفعالات همسایه است. به عبارت دیگر رویکرد محلی مقدار Read count هر فعل و انفعال را نسبت به مقدار Read count فعل و انفعالات همسایگی مورد بررسی قرار می‌دهد. در این رویکرد فرض ما بر این است که read count هر فعل و انفعال را می‌توان براساس مقدار read count فعل و انفعالات همسایه محاسبه کرد. به عبارت دیگر فرض کرده‌ایم که مقدار read count هر فعل و انفعال مشابه و نزدیک به مقدار read counts فعل و انفعالات همسایه آن است. برای این منظور ما از شبکه عصبی اتوانکدر استفاده کرده‌ایم. ورودی این شبکه مقدار read count، فعل و انفعالات است که از نظر مکانی نزدیک به هم هستند. برای این منظور ما داده‌ها را بر اساس مکان وقوع فعل و انفعالات با استفاده از K-means خوشه بندی نموده‌ایم که در نتیجه ورودی در این شبکه به صورت خوشه‌ای (خوشه‌ها مشخص کننده فعل و انفعالات نزدیک به یکدیگر می‌باشند) اعمال می‌شود.

به طور کلی روشی پیشنهادی دارای چهار مرحله اساسی است (شکل ۷-۳). در مرحله اول به مدل‌سازی دادگان با استفاده از یک مدل آماری مانند روش ارائه شده در GOTHiC پرداخته می‌شود. در ادامه فرایند، در مرحله دوم فعل و انفعالات همسایه شناسایی می‌شوند و در یک خوشه قرار می‌گیرند. در مرحله سوم با استفاده از شبکه عصبی اتوانکدر دادگان Hi-C را بر اساس مولفه‌های محلی و فعل و

انفعالات همسایه مدل می‌کنیم. در واقع در این مرحله و به نوعی مدل آماری بدست امده را بهبود می‌بخشیم. مرحله چهارم روش پیشنهادی به شناسایی فعل و انفعالات معنی‌دار پرداخته می‌شود. در ادامه این بخش هر کدام از این مراحل با جزییات بیشتر شرح داده می‌شوند.

شکل ۷-۳ چهار مرحله اساسی روش پیشنهادی

۳-۳-۲ مرحله اول (مدل‌سازی)

همان طور که در فصل اول توضیح داده شد، تعداد Read count هر فعل و انفعال به مولفه‌های مختلفی وابسته است. یکی از این مولفه‌ها، مولفه فاصله است و طبق فرض مساله تعداد Read count هر فعل و انفعال به فاصله بین دو ناحیه کروموزومی وابسته است. برای مدل کردن رابطه فعل و انفعالات و این نوع مولفه‌ها می‌توان از مدل‌های آماری استفاده کرد. تا کنون مدل‌های آماری مختلفی برای مدل‌سازی داده‌ها به خصوص در حوزه Hi-C ارائه شده‌اند. با این وجود در این پژوهش ما از مدل ارائه شده در روش GOTHiC به دلیل ماهیت داده‌ها و سرعت پردازش استفاده نموده‌ایم. در واقع روش پیشنهادی در این مرحله یک رویکرد سراسری در پیش گرفته و با استفاده از مدل ارائه شده در روش GOTHiC رابطه بین مقدار Read count و مولفه فاصله را بدست می‌آورد.

ورودی این مرحله، مجموعه‌ای از اطلاعات مورد نیاز برای شرح یک فعل و انفعال Hi-C است. این

اطلاعات مانند جدول ۱-۳ در کمترین حالت از پنج ستون locus position 1، chromosome 1، locus position 2، chromosome 2 و Read Counts تشکیل شده است. در این دادگان ستون‌های locus position و chromosome به ترتیب نوع کروموزوم و مکان هر ناحیه شرکت کننده در یک فعل و انفعال را مشخص می‌کنند. همچنین خروجی این مرحله مقدار Read Counts تخمینی بر اساس رگرسیون مورد استفاده است.

جدول ۱-۳ بخشی از دادگان Dixon استفاده شده در این پژوهش که در مرحله اول مورد استفاده قرار می‌گیرد

locus position	Chromosome	locus position	Chromosome	Read Counts
0	chromosome*1*	۱۰۰۰۰۰۰	chromosome*1*	۳۵۳
۱۴۲۰۰۰۰۰۰	chromosome*1*	۱۶۵۰۰۰۰۰۰	chromosome*1*	۵
۱۶۴۰۰۰۰۰۰	chromosome*1*	۱۶۵۰۰۰۰۰۰	chromosome*1*	۶۹۶

۳-۳-۳ مرحله دوم (خوشه‌بندی)

داشتن یک رویکرد محلی یکی از مزیت‌های روش پیشنهادی نسبت به روش‌های دیگر است. همان طور که در بخش قبل اشاره شده، در این پژوهش رویکرد محلی به معنای بررسی و پردازش دادگان Hi-C در گروهی از فعل و انفعالات همسایه است. به عبارت دیگر روش ارائه شده در این پژوهش برای شناسایی و حذف نویز علاوه بر بررسی هر فعل و انفعال، فعل و انفعالاتی که در همسایگی آن فعل و انفعال قرار دارند را نیز بررسی می‌کند. دلیل استفاده از رویکرد محلی در این پژوهش، این بوده است که فعل و انفعالات نزدیک به هم می‌بایست، مشخصات و ویژگی‌های مشابه‌ای داشته باشند. همچنین تاثیر نویز بر همه‌ی داده یکسان نبوده و در نتیجه احتمال تغییر در همه‌ی فعل و انفعالات به صورت یکسان بسیار کم است. لذا مقادیر نا هماهنگ یک فعل و انفعال نسبت به فعل و انفعالات همسایه می‌تواند نشان دهنده‌ی نویز وخطا در دادگان باشد. با توجه به این موضوع می‌توان نویز موجود در یک

فعل و انفعال را با توجه به فعل و انفعالات همسایه بررسی و شناسایی کرد.

همان طور که در بالا به آن اشاره شد، رویکرد محلی که در این پژوهش مورد استفاده قرار گرفته است، یکی از مزیت‌های روش پیشنهادی نسبت به روش‌های دیگر است. از این جهت برای داشتن یک رویکرد محلی، در این پژوهش از شبکه عصبی اتوانکدر به دلیل ساختار این شبکه استفاده کرده‌ایم. ورودی این شبکه مقدار read count فعل و انفعالاتی است که از نظر مکانی نزدیک به هم هستند. برای این منظور ما داده‌ها را بر اساس مکان وقوع فعل و انفعالات با استفاده از K-means خوشه بندی می‌کنیم. ورودی این بخش نیز همان ورودی بخش قبل است که بر اساس جدول ۱-۳ این خوشه‌بندی نسبت به ستون‌های locus position 1 و locus position 2 انجام می‌شود. همچنین خروجی این مرحله به صورت خوشه‌هایی با اندازه ثابت از فعل و انفعالات نزدیک به هم است که در مرحله بعد (شبکه عصبی اتوانکدر) مورد استفاده قرار می‌گیرند. از انجایی که ورودی شبکه عصبی در این پژوهش به صورت خوشه‌ای (خوشه‌ها مشخص کننده فعل و انفعالات نزدیک به یکدیگر می‌باشند) اعمال می‌شود. لذا می‌توان گفت مرحله دوم دادگان را برای شبکه عصبی اتوانکدر آماده می‌کند.

۳-۳-۴ مرحله سوم (اتوانکدر)

همان طور که در بخش‌های قبل به آن اشاره شد، در این پژوهش از شبکه عصبی اتوانکدر برای مدل کردن داده‌ها استفاده کرده‌ایم. ایده اصلی استفاده از شبکه عصبی بر این فرض استوار است که مقدار Read count هر فعل و انفعال می‌بایست نزدیک به مقدار مقدار Read count فعل و انفعالات همسایه‌اش باشد. لذا برای شناسایی و استفاده از ویژگی‌های محلی از شبکه عصبی اتوانکدر استفاده کرده‌ایم. به طور کلی دلایل مختلفی وجود دارد که ما شبکه عصبی اتوانکدر را برای مدل کردن دادگان انتخاب کرده‌ایم. دلایل استفاده از این شبکه عصبی در روش ارائه شده به شرح زیر است.

- به دلیل اینکه دادگان Hi-C به صورت آزمایشگاهی استخراج می‌شوند و همچنین نمونه داده نرمال شده برای این دادگان وجود ندارد. لذا برای حذف نویز و یافتن الگوهای پنهان در این

دادگان می‌بایست از روش‌های یادگیری ماشینی بدون نظارت استفاده کرد. از آنجایی که شبکه عصبی اتوانکدر می‌تواند یادگیری بدون نظارت داشته باشد، لذا می‌توان از این روش بدون هیچ مشکلی در مدل کردن دادگان استفاده نمود.

- همان طور که در فصل قبل به آن اشاره شد، شبکه عصبی اتوانکدر برای استخراج ویژگی از دادگان و بازسازی دادگان بر اساس این ویژگی‌ها، دارای ساختار دو مرحله‌ای است. مرحله اول این شبکه، ویژگی‌های مهم دادگان را به صورت خودکار استخراج می‌کند. در مرحله دوم، این شبکه سعی می‌شود که بر اساس ویژگی‌های استخراج شده دادگان را بازسازی نماید. از آنجایی که تاثیر نویز بر ویژگی‌های استخراج شده کم است، لذا بازسازی دادگان بر اساس ویژگی‌های استخراج شده می‌تواند تا حد زیادی خطا و نویز موجود در دادگان را حذف نماید. این مساله مهمترین دلیلی است که در این پژوهش برای حذف نویز از شبکه عصبی اتوانکدر استفاده نموده‌ایم.

در این پژوهش برای استفاده از شبکه اتوانکدر و همچنین برای بدست آوردن مدلی بهتر، ساختار دادگان عوض شده است. همان طور که در بخش قبل توضیح داده شد برای استفاده از این شبکه عصبی ابتدا دادگان خوشه بندی شده و به صورت خوشه‌هایی از فعل و انفعالات همسایه (فعل و انفعالاتی که از نظر مکان قرارگیریشان نزدیک به هم می‌باشد.) به شبکه عصبی اتوانکدر اعمال می‌شوند. از آنجایی که ورودی شبکه به صورت خوشه‌ای است. لذا تعداد زیادی خوشه برای آموزش شبکه عصبی و پیدا کردن روابط بین فعل و انفعالات همسایه مورد استفاده قرار می‌گیرند. همچنین تعداد ورودی‌های شبکه برابر با تعداد فعل و انفعالات موجود در خوشه‌ها است. به عبارت دیگر مقدار read count هر فعل و انفعال به صورت یک ورودی مجزا به شبکه عصبی اعمال می‌شود که تعداد ورودی‌های این شبکه برابر با اندازه هر خوشه است. در نهایت این شبکه بر اساس مقدار ورودی اولیه سعی می‌کند که رابطه بین ورودی‌ها (تاثیر فعل و انفعالات نزدیک بر یکدیگر) را مدل کند و مقدار ورودی را بر اساس تاثیرات فعل و انفعالات نزدیک به یکدیگر بازسازی نماید. به طور کلی در این پژوهش ما از شبکه عصبی اتوانکدر در

یک مسئله رگرسیون استفاده کرده‌ایم که در آن شبکه عصبی اتونکدر مدلی را بر اساس ویژگی‌های محلی دادگان ارائه می‌دهد. به عبارت دیگر بعد از خوشه‌بندی دادگان، اگر مجموعه X با اعضای $\{x_1, x_2, \dots, x_n\}$ را مقادیر اولیه read counts و مجموعه Y شامل اعضای $\{Y_1, Y_2, \dots, Y_n\}$ را مقادیر تخمینی read counts در یک خوشه در نظر بگیریم. شبکه عصبی اتوانکدر سعی می‌کند رابطه‌ی بین مجموعه مقادیر اولیه read counts یک خوشه (X) و مقدار تخمینی هر read count (Y_i) را مدل می‌کند. در نتیجه پس از آموزش شبکه عصبی و بدست آوردن مدل، مقدار read counts را می‌توان بر اساس مدل بدست آمده تخمین زد. در نتیجه با فرض اینکه y نشان دهنده مقدار تخمینی جدید باشد، می‌توان نتیجه گرفت:

$$y_i = f(X, Y_i) \qquad (۳-۶)$$

۳-۳-۵ مرحله چهارم (شناسایی فعل و انفعالات معنی‌دار)

بعد از محاسبه مقدار تخمینی read count برای هر فعل و انفعال در مرحله قبل، در این پژوهش فرایند شناسایی فعل و انفعالات معنی‌دار در داده‌های Hi-C شروع می‌شود. فعل و انفعالات معنی‌دار در واقع فعل و انفعالاتی هستند که از نظر آماری معنادار شناسایی شده‌اند. فعل و انفعالات معنی‌دار نشان می‌دهند که تا چه میزان مقادیر تخمین زده شده به صورت تصادفی ایجاد نشده‌اند. به طور کلی می‌توان گفت شناسایی فعل و انفعالات معنی‌دار در واقع ارزیابی نتیجه بدست آمده در مرحله قبل است. به عبارت دیگر در این مرحله با استفاده از این مفهوم، درستی مقدار تخمینی read count هر فعل و انفعال مورد ارزیابی قرار می‌گیرد. همچنین از آنجایی که فعل و انفعالات معنی‌دار، فعل و انفعالاتی هستند که مقدار Read count بیشتری نسبت به بقیه فعل و انفعالات دارند. لذا بر اساس این فعل و انفعالات می‌توان نواحی که کمترین فاصله را دارند و به طور کلی اهمیت بالای در ساخت پیکربندی کروموزوم دارند را مشخص نمود. در نتیجه پیکربندی کروموزوم را می‌توان با تعداد کمتری داده پیش‌بینی کرد و بار محاساباتی را کم نمود.

در شناسایی فعل و انفعالات معنی‌دار در این پژوهش از آزمون دوجمله‌ای استفاده کرده‌ایم. برای این منظور ابتدا پس از بدست آمدن مقدار تخمینی read count، احتمال هر یک از فعل و انفعالات را بر اساس مقادیر تخمین زده شده جدید محاسبه می‌شوند. این احتمالات که مشخص کننده حدس ما از مقدار read count است به عنوان فرض صفر مساله در نظر گرفته می‌شوند. در نتیجه بر اساس این احتمالات و مقدار اولیه read count و با استفاده از آزمون دوجمله‌ای پیرو روش Benjamini–Hochberg multiple testing correction، مقادیر read count هر فعل و انفعال مورد ارزیابی قرار می‌گیرند. به عبارت دیگر بر اساس این روش مقادیر p-value و به تبع آن q-value (adjusted p-value) را برای هر فعل و انفعال محاسبه می‌شوند. حال با استفاده از مقادیر q-value می‌توان مشخص کرد که کدام فعل و انفعالات معنی‌دار هستند. برای مشخص کردن فعل و انفعالات معنی‌دار از مقادیر q-value، نیاز است که حد آستانه‌ای (سطح معناداری) برای این مقادیر مشخص شود. برای توضیح بهتر این موضوع می‌توان مسئله را به صورت ساده در نظر گرفت. در این حالت اگر مقدار اولیه Read count هر فعل و انفعال بسیار کوچکتر از مقدار تخمین زده شده باشد، فعل و انفعال مورد نظر معتبر نبوده است و در نتیجه مقدار تخمینی زده شده امکان پذیر نبوده است.

۳-۴ خلاصه مطالب

در این فصل روش توسعه داده شده برای حذف خطاهای سیستماتیک همراه با ابزار تولید شده برای این منظور شرح داده شده‌اند. برای ابزار طراحی شده ما چندین روش را همراه با توانایی دریافت داده‌های مختلف پیاده‌سازی کرده‌ایم. همچنین ویژگی‌هایی مانند بصری‌سازی داده‌های Hi-C به صورت معنی‌دار را در آن توسعه داده‌ایم. در قسمت دیگر این فصل روش پیشنهادی شرح داده شده است. دراین روش دو رویکرد متفاوت در شناسایی نویز را ادغام کرده‌ایم. در واقع ما با در نظر گرفتن یک رویکرد محلی، مدل آماری بدست آمده از رویکرد سراسری را بهبود بخشیده‌ایم. به طور کلی در روش پیشنهادی ابتدا

داده‌ها با استفاده از روش GOTHiC مدل می‌شوند. سپس با استفاده از یک شبکه عصبی مصنوعی به نام اتوانکدر مدل بدست آمده را بهبود بخشیده و مقدار Read counts را بر اساس مدل بهبود یافته محاسبه می‌کند. در پایان هم فعل و انفعالات معنی‌دار (معتبر) به صورت آماری شناسایی می‌شوند.

فصل ۴ تجزیه و تحلیل نتایج پژوهش

همان طور که در فصل‌های گذشته شرح داده شده، داده‌های Hi-C به صورت آزمایشگاهی تولید می‌شوند. در فرایند استخراج این دادگان عموما خطاهای سیستماتیک رخ می‌دهند. لذا برای حذف این خطاها روش‌های مختلفی از جمله روش پیشنهادی در این رساله توسعه داده شدند. از آنجایی که نمونه داده بدون خطا برای این موضوع وجود ندارد، برای ارزیابی روش توسعه داده شده نسبت به روش‌های دیگر نمی‌توان از مقایسه خروجی بدست آمده با داده بدون خطا استفاده کرد. لذا برای ارزیابی روش پیشنهادی در این رساله از شناسایی فعل و انفعلات معنی‌دار و ضریب همبستگی پیرسون استفاده کرده‌ایم. به طور کلی در این فصل ابتدا به شرح محیط ارزیابی پرداخته می‌شود و سپس در ادامه فصل پیاده سازی انجام شده شرح داده می‌شود و بعد از آن نتایج حاصل از روش پیشنهادی همراه با نتایج روش‌های GOTHiC و Fit-Hi-C آورده شده است.

۴-۱ محیط ارزیابی

همان طور که گفته شد، برای ارزیابی روش پیشنهادی در این رساله از شناسایی فعل و انفعلات معنی‌دار و ضریب همبستگی پیرسون استفاده کرده‌ایم. فعل و انفعلات معنی‌دار در واقع بیانگر غیر تصادفی بودن مقادیر محاسبه شده می‌باشد. لذا تعداد فعل و انفعلات معنی‌دار به ما نشان می‌دهد که تا چه میزان احتمالات محاسبه شده و مقادیر تخمینی صحیح می‌باشند. به عبارت دیگر هر فعل و انفعال معنی‌دار نشان دهنده این است که مقدار Read Counts تخمینی برای آن فعل و انفعال از نظر آماری غیر تصادفی بوده و صحیح می‌باشد. لذا تعداد هر فعل و انفعالات معنی‌دار نشان می‌دهد که روش پیشنهادی به درستی توانسته است دادگان را مدل کند. همچنین تعداد فعل و انفعلات معنی‌دار مشترک با روش‌های دیگر معیاری برای اندازه گیری جامع بودن آن روش نسبت به روش‌های دیگر است. این موضوع نشان می‌دهد که روش مورد بحث توانسته است به خوبی فرض‌های روش‌های دیگر را در فرایند مدل‌سازی لحاظ نماید و خروجی مشابه با روش‌های دیگر تولید نماید.

برای ارزیابی روش پیشنهادی به غیر از تعداد فعل و انفعلات معنی‌دار می‌توان از مقادیر میانگین Read Counts نیز استفاده کرد. این مقادیر می‌توانند پایبندی مدل بدست آمده به فرض مساله را مورد ارزیابی قرار دهند. به عبارت دیگر هرچه میانگین مقدار Read Counts بیشتر باشد، نشان می‌دهد تا چه میزان روش پیشنهادی توانسته است که فعل و انفعالات مناسبی را شناسایی کند.

علاوه‌بر این بررسی همبستگی بین معیار فاصله و مقدار تخمین زده شده Read counts در روش مورد بحث، پایبندی آن روش را نسبت به فرض اصلی مساله نشان می‌دهد. برای این منظور در این پژوهش برای ارزیابی این موضوع از ضریب همبستگی پیرسون استفاده شده است. ضریب همبستگی پیرسون مقدار همبستگی بین دو متغیر تصادفی را بررسی می‌کند. در این ضریب مقدار بین ۱- تا ۱ تغییر می‌کند. در این ضریب مقدار «۱» به معنای همبستگی مثبت کامل و مقادیر «۰» و «۱-» به ترتیب به معنی نبود همبستگی و همبستگی منفی کامل هستند. از آنجایی که طبق فرض مساله در این داده‌ها باید همبستگی بین مقادیر Read counts و فاصله بین دو ناحیه زیاد باشد، لذا طبق این فرض بهتر است که مقدار ضریب همبستگی پیرسون بین معیار فاصله و مقدار تخمینی Read counts به ۱ و ۱- نزدیک شوند. برای شرح بیشتر این موضوع می‌توان گفت که اگر یک روش دادگان را مدل نماید و ضریب همبستگی محاسبه شده در آن برابر با صفر باشد یا به عبارت دیگر این دو معیار فاصله و مقدار Read counts را مستقل از هم شناسایی کند، آن روش نتوانسته است به فرض اصلی مساله پایبند بوده باشد. البته قابل ذکر است که این روش ارزیابی به تنهایی نمی‌تواند قدرت یک روش را نشان دهد و فقط می‌تواند پایبندی یک روش را نسبت به فرض مساله مورد ارزیابی قرار دهد.

۴-۲ دادگان

برای ارزیابی این روش نسبت به روش‌های GOTHiC و Fit-Hi-C از دادگان IMR90 رده سلولی انسانی Dixon استفاده کرده‌ایم. در ارزیابی روش پیشنهادی به دلیل حجیم بودن و بار محاسباتی این داده‌گان فقط از کروموزوم ۱ پایگاه‌داده Dixon استفاده کرده‌ایم [41] ,[40]. همچنین این دادگان

ابتدا توسط HiC-Pro پردازش شده و اندازه هر ناحیه برابر با یک میلیون جفت باز[1] (1Mbp) در نظر گرفته شده است. به عبارت دیگر هر ناحیه شرکت کننده در یک فعل و انفعال خود دارای طولی برابر با 1Mbp است. اطلاعات مربوط به این پایگاه داده در جدول ۱-۴ آورده شده است.

جدول ۱-۴ پایگاه‌داده

کروموزوم ۱ پایگاه‌داده Dixon	
تعداد فعل و انفعالات	۲۶۳۰۲
مجموع مقادیر Read Counts	۱۰۶۵۰۶۹
میانگین مقادیر Read Counts	۴۰/۴۹
اندازه هر ناحیه	یک میلیون جفت باز

۴-۳ پیاده‌سازی

همان طور که در فصل‌های گذشته شرح داده شد، در این پژوهش علاوه‌بر روشی برای حذف نویز، ابزاری جامعی با نام MHiC برای حذف نویز و مصورسازی دادگان توسعه داده‌ایم. در این بخش پیاده‌سازی انجام گرفته برای روش و ابزار MHiC شرح داده می‌شود.

۴-۳-۱ ابزار MHiC

ابزار MHiC به طور کلی دارای دو بخش محاسبات و مصورسازی است. بخش محاسبات این ابزار شامل مراحل هماهنگ سازی دادگان ورودی، نرمال‌سازی دادگان و شناسایی فعل و انفعالات معنی‌دار است که تمامی این مراحل در مخیط R توسعه داده‌ایم. بخش دیگر ابزار MHiC، بخش مصورسازی دادگان Hi-C است. این بخش شامل نمایش دادگان به صورت Contact map diagram و Arc diagram است که تمامی مراحل موجود در این بخش را با استفاده از Javascript توسعه داده‌ایم. همچنین برای سادگی استفاده از ابزار MHiC، این بخش را با استفاده از پکیج Shiny server به محیط R متصل

[1] Base pair

نموده‌ایم. در واقع در بخش مصورسازی دادگان علاوه‌بر نمایش دادگان، رابط کاربری گرافیکی نیز برای بخش محاسبات ایجاد کرده‌ایم. لازم به ذکر است که بخش محاسبات ابزار MHiC به صورت یک پکیج قابل نصب در محیط R و بخش مصورسازی همراه با بخش محاسبات در آدرس github.com/MHi-C در دسترس عموم قرار دارند.

۴-۳-۲ حذف نویز در MHiC

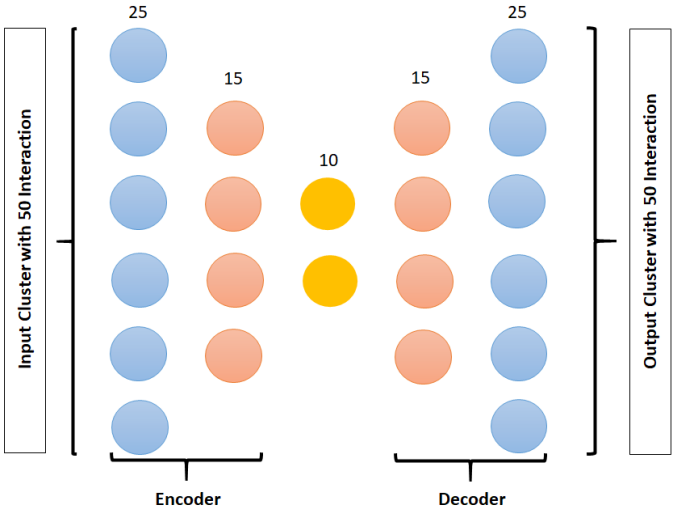
در این پژوهش، روش MHiC را با سه شیوه متفاوت در بخش مدل‌سازی آماری، در محیط R توسعه داده‌ایم. ابتدا از رگرسیون دوجمله‌ای منفی برای مدل کردن داده‌های Hi-C در روش MHiC (MHiC v1) استفاده کرده‌ایم. سپس برای بهبود نتایج MHiC v1 از روش ارائه شده در GOTHiC برای تولید مدل آماری اولیه از دادگان، استفاده کرده‌ایم (MHiC v2). همچنین برای درک تاثیر مدل آماری در کل فرایند مدل‌سازی و حذف نویز، روش MHiC را بدون استفاده از مدل‌سازی آماری نیز پیاده سازی نموده‌ایم (MHiC v3). لازم به ذکر است که در پیاده‌سازی انجام گرفته، در تمامی این روش‌ها از خوشه‌بندی K-means استفاده شده است و همچنین برای آموزش شبکه عصبی (به دلیل ثابت بودن اندازه ورودی) اندازه هر خوشه برای دادگان استفاده شده در ارزیابی را برابر با پنجاه فعل و انفعال در نظر گرفته‌ایم. در این پژوهش اندازه هر خوشه به صورت تجربی براساس تعداد خوشه‌های تولیدی و نتیجه حاصل از شبکه عصبی اتوانکدر تعیین گردیده است. بر اساس آزمایش‌های انجام شده زمانی که اندازه خوشه‌ها کوچک در نظر گرفته شود، ساختار شبکه عصبی اتوانکدر کوچک می‌شود. به عبارت دیگر از تعداد لایه و تعداد نورون‌های داخل هر لایه کاسته می‌شود. لذا این شبکه عصبی با ساختار ایجاد شده نمی‌تواند مدلی از دادگان را تولید نماید. لازم به ذکر است که در این حالت نیازی به استفاده از خوشه‌بندی برای تعیین فعل و انفعالات همسایه نیست. همچنین زمانی که اندازه خوشه‌ها بزرگتر از پنجاه در نظر گرفته شده بودند، به دلیل تعداد محدود فعل و انفعالات موجود در دادگان Dixon خوشه‌های کمتری تولید شده بودند. لذا با تعداد خوشه کمتر نمی‌توان شبکه عصبی بزرگتر را آموزش

داد و دادگان را مدل نمود.

علاوه بر بخش خوشه‌بندی در پیاده‌سازی انجام گرفته از یک شبکه عصبی اتوانکدر پنج لایه مطابق شکل ۱-۴ استفاده نموده‌ایم. تعداد لایه‌ها و تعداد نورون‌های این شبکه عصبی نیز مانند اندازه هر خوشه بر اساس دادگان Dixon به صورت تجربی بدست آمده است. در واقع ابتدا ما شبکه عصبی را با سه لایه و سپس با ۵ لایه و ۷ لایه پیاده‌سازی نموده‌ایم که بر اساس نتایج بدست آمده برای روش MHiC v2 در این آزمایش ما از شبکه عصبی اتوانکدر ۵ لایه استفاده کرده‌ایم. نتایج بدست آمده از ساختارهای مختلف شبکه عصبی در جدول ۲-۴ قابل مشاهده است.

جدول ۲-۴ نتایج بدست آمده از ساختارهای مختلف شبکه عصبی اتوانکدر با استفاده از روش MHiC v2

ساختار شبکه عصبی اتوانکدر	تعداد فعل و انفعالات معنی‌دار
۳ لایه	۳۱۹۷
۵ لایه	۳۷۷۷
۷ لایه	۳۳۲۵



شکل ۱-۴ ساختار شبکه عصبی استفاده شده در این پژوهش

به طور کلی تمامی روش‌های پیاده‌سازی شده فقط در بخش مدل‌سازی آماری تفاوت دارند و در بخش‌های دیگر ساختار یکسانی دارند. در ادامه این فصل نتایج مربوط به پیاده‌سازی روش‌های MHiC v1، MHiC v2 و MHiC v3 و همچنین نتایج حاصل از روش‌های GOTHiC و Fit-Hi-C آورده شده است.

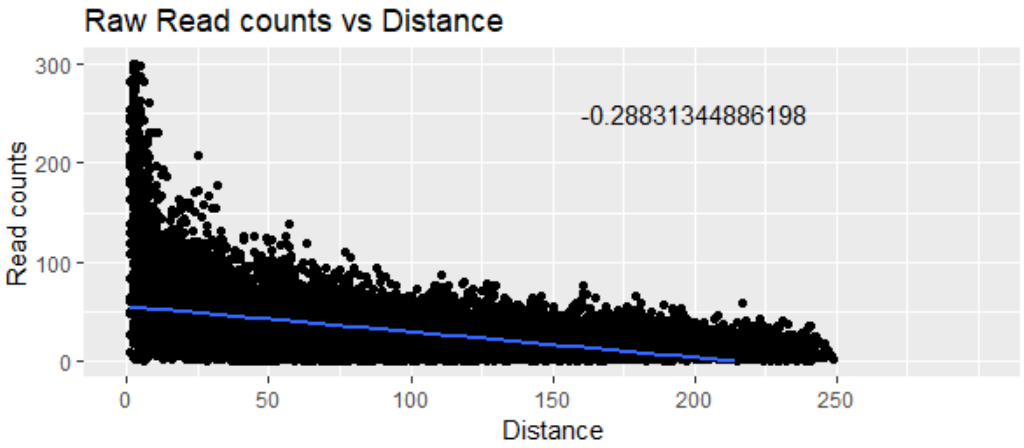
۴-۴ نتایج

برای ارزیابی روش پیشنهادی و مقایسه آن با روش‌های GOTHiC و Fit-Hi-C، از دادگان معرفی شده در بخش قبل همراه با روش‌های ذکر شده در بخش محیط ارزیابی استفاده کرده‌ایم.

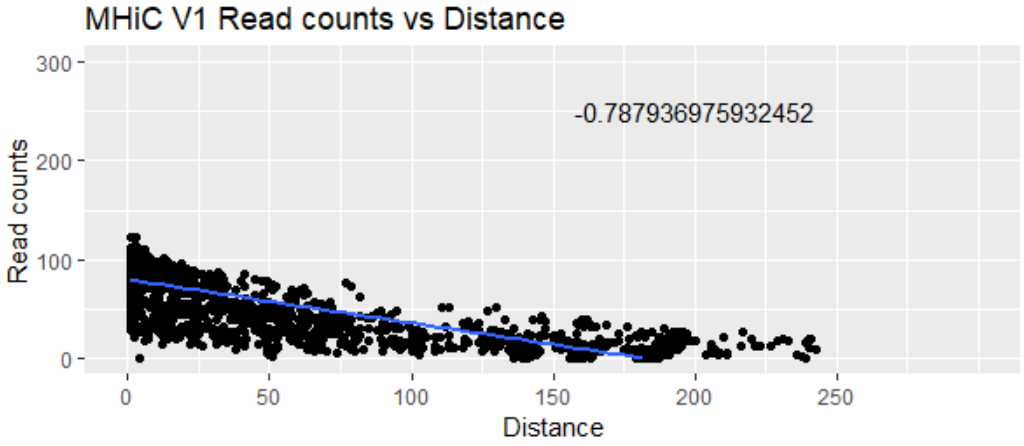
برای ارزیابی، ابتدا پایبندی روش پیشنهادی را نسبت به فرض اصلی مساله مورد ارزیابی قرار داده‌ایم. در بررسی این موضوع، برای پیدا کردن همبستگی بین معیار فاصله و مقادیر تخمین زده شده روش‌های مختلف و همچنین مقادیر خام، ما از ضریب همبستگی پیرسون استفاده کرده‌ایم. برای محاسبه این ضریب همبستگی از فرمول ۴-۱ استفاده کرده‌ایم. در این معادله x و y نشان دهنده مولفه‌هایی است که مورد ارزیابی قرار می‌گیرند که این مولفه‌ها در این رساله همان مقدار Read counts و معیار فاصله کروموزومی می‌باشد.

$$p_{x,y} = \frac{cov(x,y)}{\sigma_x . \sigma_y} \qquad (۱-۴)$$

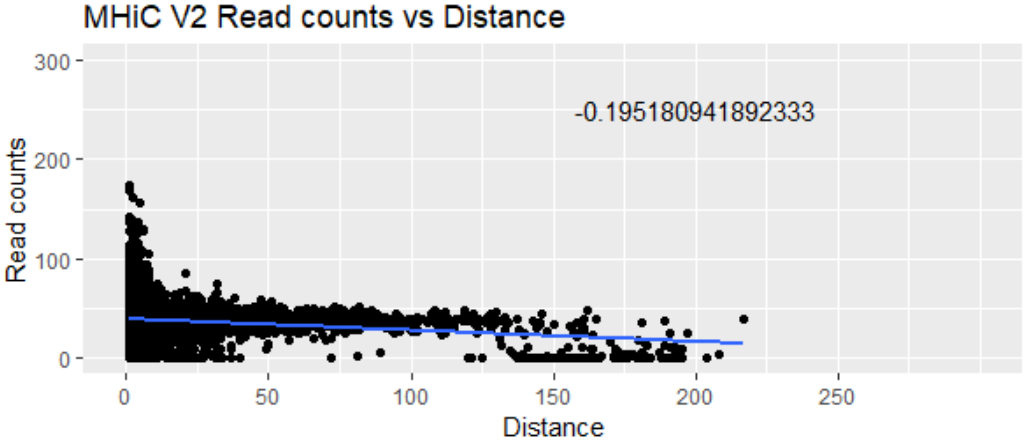
نتایج حاصل از این آزمایش در شکل ۴-۲ تا شکل ۴-۷ ارائه شده است. لازم به ذکر است که این نمودارها نشان دهنده توزیع داده‌های Hi-C نسبت به فاصله کروموزومی و همچنین مقدار ضریب همبستگی می‌باشند.



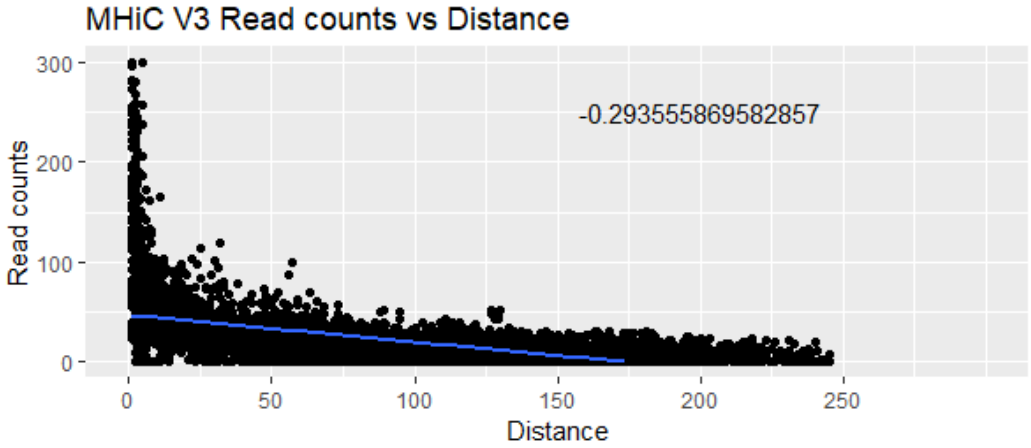
شکل ۲-۴ مقدار ضریب همبستگی پیرسون بین معیار فاصله و مقدار اولیه Read counts



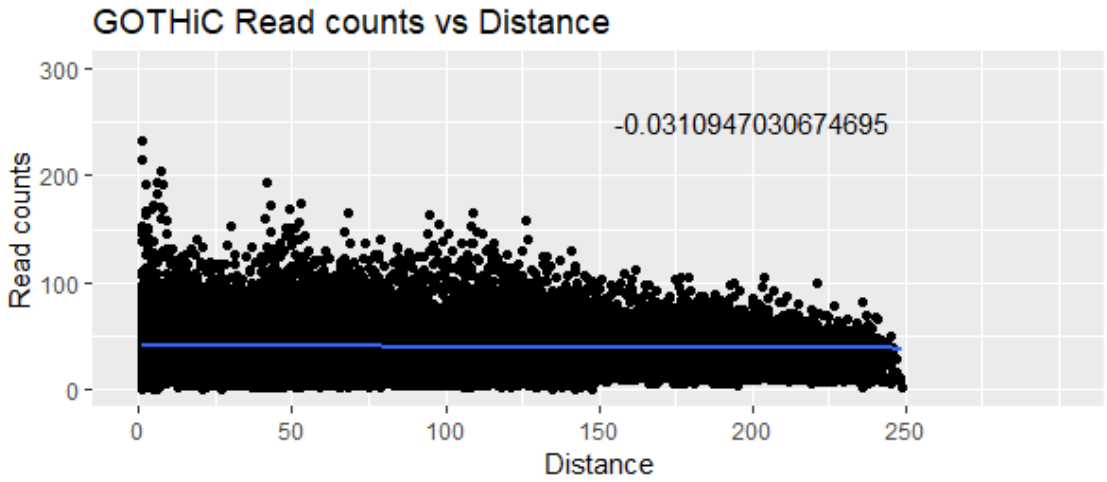
شکل ۳-۴ مقدار ضریب همبستگی برای روش پیشنهادی با استفاده از روش رگرسیون دو جمله‌ای



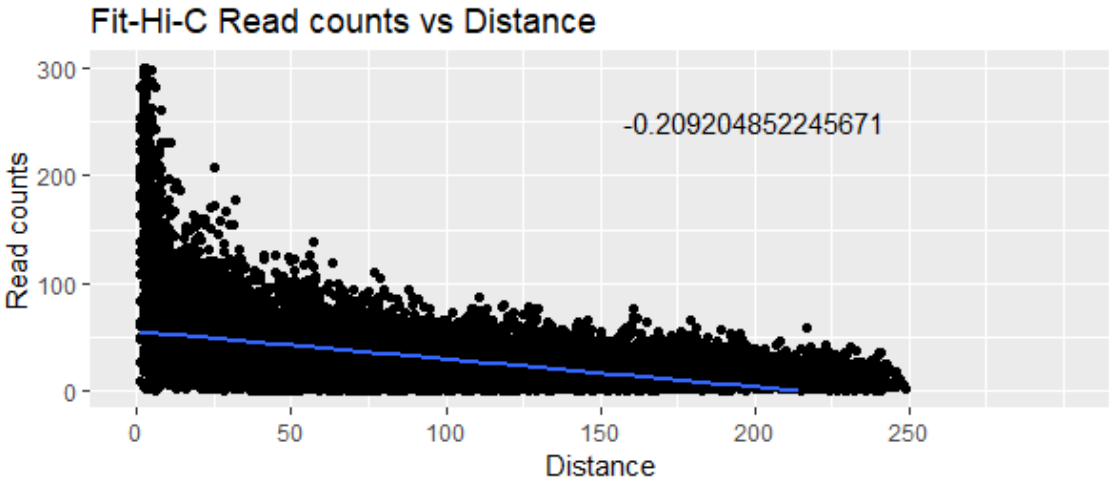
شکل ۴-۴ مقدار ضریب همبستگی برای روش پیشنهادی با استفاده از روش GOTHiC



شکل ۴-۵ مقدار ضریب همبستگی برای روش پیشنهادی بدون استفاده از مدل اولیه



شکل ۴-۶ مقدار ضریب همبستگی برای روش GOTHiC



شکل ۴-۷مقدار ضریب همبستگی برای روش Fit-Hi-C

همانطور که در شکل ۴-۲ دیده می‌شود، مقدار ضریب همبستگی بین معیار فاصله و مقدار اولیه Read counts در داده‌های مذکور برابر با ۰/۲۸۸۳- بدست آمده است. این مقدار به ما نشان می‌دهد که بین

معیار فاصله و مقدار اولیه Read counts نیز وابستگی وجود دارد. همچنین در شکل ۶-۴ و شکل ۷-۴ مشاهده می‌شود که مقدار ضریب همبستگی برای روش‌های Fit-Hi-C و GOTHiC به ترتیب با ۰/۲۰۹۲- و ۰/۰۳۱۰- برابر است. لازم به ذکر است که طبق شکل ۳-۴ تا شکل ۵-۴ مقدار این ضریب برای روش‌های MHiC v1، MHiC v2 و MHiC v3 به ترتیب با ۰/۷۸۷۹- ، ۰/۱۹۵۱- و ۰/۲۹۳۵- برابر است. لذا همان طور که در مقادیر ضرایب همبستگی مشاهده می‌شود، در این دادگان روش GOTHiC به دلیل اینکه از معیار فاصله به طور مستقیم استفاده نکرده است، ضریب همبستگی آن به صفر (وابستگی بین دو مولفه کاهش می‌یابد) نزدیک می‌شود. از این رو روش GOTHiC نتوانسته است وابستگی بین معیار فاصله و Read counts را به خوبی مدل کند. همچنین بر اساس مقدار ضریب همبستگی روش Fit-Hi-C مشاهده می‌شود که روش Fit-Hi-C توانسته است دادگان را به خوبی مدل نماید و به فرض مساله پایبند باشد. با این وجود ضریب همبستگی محاسبه شده در این روش نسبت به ضریب همبستگی بین معیار فاصله و مقدار اولیه Read counts کمتر شده است. لذا فرایند مدل‌سازی این روش به فرض اصلی مساله کاملا پایبند نبوده است.

در نهایت همان طور که در شکل ۳-۴ تا شکل ۵-۴ مشاهده می‌شود که در روش پیشنهادی زمانی که از رگرسیون دوجمله‌ای منفی (معیار فاصله مستقیم به در مدل کردن دادگان استفاده شده است) برای ایجاد مدل اولیه استفاده شده است، این روش توانسته است به خوبی وابستگی بین معیار فاصله و Read counts در دادگان را مدل نماید و بهترین نتیجه را داشته باشد. همچنین در نتایج بدست آمده مشاهده می‌شود، زمانی که روش پیشنهادی در مرحله اول فرایند حذف نویز از روش GOTHiC استفاده کرده، نتیجه به مراتب بهتری نسبت به روش ارائه شده در GOTHiC بدست آورده است. به طور کلی در حالت‌های مختلف، روش پیشنهادی به خوبی وابستگی بین معیار فاصله و Read counts در دادگان را مدل می‌کند. همچنین روش پیشنهادی در پیاده‌سازی‌های مختلف به فرض اصلی مساله نسبت به روش‌های GOTHiC و Fit-Hi-C پایبندی بیشتری داشته است.

همان طور که در بخش محیط ارزیابی توضیح داده شد، برای بررسی درست بودن مقادیر تخمین زده

شده می‌توان از شناسایی فعل و انفعلات معنی‌دار استفاده کرد. لذا برای ارزیابی مقادیر تخمینی روش پیشنهادی، در این پژوهش از این روش ارزیابی استفاده کرده‌ایم. تعداد فعل و انفعلات معنی‌دار در واقع معیاری برای اندازه‌گیری درستی مدل را به ما ارائه می‌دهد. برای شناسایی فعل و انفعلات معنی‌دار در این پژوهش از آزمون دوجمله‌ای پیرو روش Benjamini–Hochberg multiple testing correction با حد آستانه برابر ۰/۰۵ استفاده کرده‌ایم. بر اساس این حد آستانه نتایج این آزمایش در جدول ۳-۴ گزارش شده است.

جدول ۳-۴ تعداد فعل و انفعلات معنی‌دار شناسایی شده

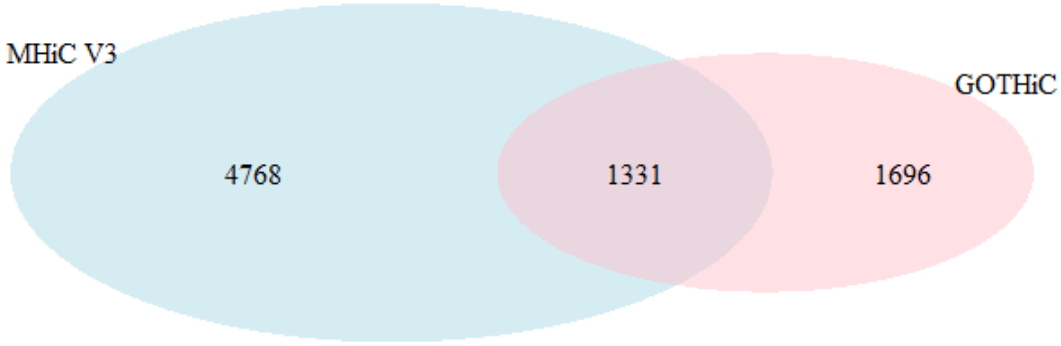
	MHiC v1	MHiC v2	MHiC v3	GOTHiC	Fit-Hi-C
تعداد فعل و انفعالات معنی‌دار	۱۹۴۷	۳۷۷۷	۶۰۹۹	۳۰۲۷	۷۲۵۷
میانگین مقادیر Read Counts	۵۸/۴۴	۳۶/۶۲	۳۵/۱۷	۴۱/۴۳	۱۱/۰۷
تعداد کل فعل و انفعالات	۲۶۳۰۲	۲۶۳۰۲	۲۶۳۰۲	۲۶۳۰۲	۶۳۸۹۷

طبق نتایج حاصل شده از این آزمایش (ارائه شده در جدول ۳-۴) می‌توان مشاهده کرد، در حالتی که روش پیشنهادی برای مدل‌سازی از روش GOTHiC استفاده می‌کند (MHiC v2) و همچنین زمانی که از هیچ روشی برای مدل کردن اولیه دادگان استفاده نمی‌کند (MHiC v3)، نسبت به روش GOTHiC فعل و انفعلات معنی‌دار بیشتری به صورت آماری شناسایی شده است. در نتیجه مقادیر تخمین زده شده در این روش نسبت به روش GOTHiC اعتبار بالاتری دارند. به عبارت دیگر در روش پیشنهادی تعداد بیشتری فعل و انفعال وجود دارد که مقدار تخمینی Read count آنها طبق آزمون دوجمله‌ای معتبر (معنی‌دار) شناسایی شده است. با این وجود روش پیشنهادی نسبت به روش Fit-Hi-C تعداد فعل و انفعالات کمتری دارد. دلیل این تفاوت در فرایند مدل‌سازی روش Fit-Hi-C نهفته است. در روش Fit-Hi-C بر خلاف روش GOTHiC و روش‌های MHiC ناحیه‌ها اندازه ثابت و یکسانی ندارند در نتیجه اندازه برخی از ناحیه در این روش کوچکتر از 1Mbp است. لذا تعداد فعل و انفعالات تولید

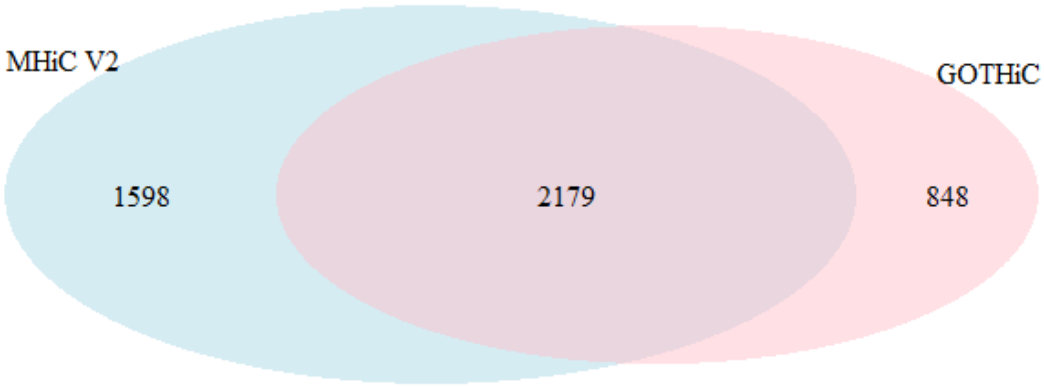
شده در این روش بسیار بیشتر بوده است و نسبت به تعداد کل فعل و انفعالات روش MHiC v3 عملکردی مشابه و تا حدودی بهتری نسبت به روش Fit-Hi-C داشته است.

همچنین در روش MHiC میانگین مقادیر Read Counts بیشتری در فعل و انفعلات معنی‌دار شناسایی شده نسبت به روش Fit-Hi-C دارد. از این جهت می‌توان گفت که روش ما فعل و انفعالات معنی‌دار بهتری را شناسایی کرده است. لازم به ذکر است که در این پژوهش فعل و انفعالات مشترک بین روش‌های MHiC و GOTHiC نیز محاسبه شده است که نتایج آن در شکل ۸-۴ تا شکل ۱۰-۴ قابل مشاهده است. هدف از طرح این مساله، نشان دادن جامعیت و قدرت روش پیشنهادی بوده است. به عبارت دیگر در این نمودارها مشاهده می‌شود تا چه میزان از فعل و انفعالات معنی‌دار شناسایی شده توسط روش GOTHiC توسط روش‌های MHiC قابل باز تولید می‌باشد. در واقع این مقدار نشان می‌دهد که تا چه میزان از مزایای روش GOTHiC در این روش استفاده شده است. همان طور که در شکل ۸-۴ تا شکل ۱۰-۴ مشاهده می‌شود، با وجود ماهیت متفاوت روش MHiC تمامی روش‌های ارائه شده توانسته‌اند تعداد قابل توجهی از فعل و انفعالات معنی‌دار شناسایی شده توسط روش GOTHiC را باز تولید نمایند.

شکل ۸-۴ تعداد فعل و انفعالات مشترک بین روش MHiC v1 و و روش GOTHiC



شکل ۹-۴ تعداد فعل و انفعالات مشترک بین روش MHiC v3 و روش GOTHiC



شکل ۱۰-۴ تعداد فعل و انفعالات مشترک بین روش MHiC v2 و روش GOTHiC

در این پژوهش روش پیشنهادی را با سه تغییر اساسی در مرحله اول فرایند حذف نویز پیاده‌سازی

نموده‌ایم (روش‌های MHiC v1، MHiC v2 و MHiC v3). همان طور که مشاهده می‌شود زمانی که از رگرسیون دو جمله‌ای منفی استفاده شده است، مدل بدست آمده کاملا به فرض اول مساله پایبند بوده است. با این وجود این روش به دلیل روش مدل سازی اولیه نتوانسته است دادگان را به خوبی مدل نماید. لذا برای رفع این مشکل ما بخشی از روش GOTHiC را برای ایجاد مدل اولیه استفاده کرده‌ایم. همچنین همان طور که گفته شد روش پیشنهادی را بدون مدل‌سازی اولیه نیز پیاده‌سازی کرده‌ایم. با توجه به نتایجی که از دو پیاده‌سازی اخیر روش MHiC حاصل شده است.

طبق نتایج بدست آمده مشاهده می‌شود که روش ارائه شده در این پژوهش نسبت به روش GOTHiC، تعداد فعل و انفعلات معنی‌دار بیشتری را شناسایی کرده است و همچنین روش پیشنهادی توانسته است تعدا قابل توجهی از فعل و انفعالات شناسایی شده در روش GOTHiC را نیز شناسایی کند (شکل ۹-۴ و شکل ۱۰-۴). در ضمن با توجه به ضریب همبستگی پیرسون، این روش به خوبی توانسته است که دادگان را طبق فرض اول مساله مدل نماید. لذا روش پیشنهادی در دو پیاده سازی اخیر عملکردی به مراتب بهتری نسبت به روش GOTHiC در مدل‌سازی و حذف نویز از دادگان داشته است. همچنین با مقایسه نتایج روش پیشنهادی با روش Fit-Hi-C مشاهده می‌شود که این روش پیشنهادی عملکرد برابر و تا حدودی بهتری نسبت به روش Fit-Hi-C داشته است.

۴-۵ خلاصه مطالب

در این فصل ما به ارزیابی روش توسعه داده شده در این پژوهش پرداختیم و آن را با روش‌های GOTHiC و Fit-Hi-C مقایسه نموده‌ایم. با ارزیابی ضریب همبستگی پیرسون بین معیار فاصله و مقادیر تخمین زده شده، می‌توان نتیجه گرفت که این روش به خوبی توانسته است داده‌های Hi-C را بر اساس فرضیات اولیه و اصلی مساله مدل کند و همچنین عملکرد به مراتب بهتری را نسبت به روش‌های دیگر داشته باشد. با ارزیابی تعداد فعل و انفعالات معنی‌دار به این نتیجه رسیدیم که روش ما نسبت به روش GOTHiC عملکرد به مراتب بهتری داشته است و همچنین نسبت به روش Fit-Hi-C عملکرد مشابه‌ای

داشته است. به طور کلی در این فصل می‌توان گفت که روش ارائه شده در این پژوهش از جهات مختلف عملکرد خوب و قابل قبولی برای حذف نویز داشته است.

فصل ۵ نتیجه گیری و سوی کارهای آتی

۵-۱ خلاصه تحقیق

در این پژوهش هدف ما ارائه یک روش قابل اعتماد برای حذف نویز و شناسایی فعل و انفعالات معنی‌دار از داده‌های Hi-C بوده است. پروتکل Hi-C یک روش مبتنی بر پروتکل 3C است. این پروتکل برای بالا بردن بازده استخراج اطلاعات ساختاری کروموزوم‌ها، توسعه داده شده است. از انجایی که Hi-C یک پروتکل آزمایشگاهی هست، این داده‌ها دارای خطاهای سیستماتیک و نویز زیادی است. از آنجایی که برای تحلیل این داده‌ها نیاز به نرمال‌سازی این داده‌ها می‌باشد، تا کنون روش‌های زیادی بر اساس مدل‌های آماری برای حذف نویز توسعه یافته‌اند. هر یک از روش‌های توسعه یافته دارای مشکلات خاص خود می‌باشد. بنابراین در این پژوهش ما سعی کردیم تا حد امکان این مشکلات را حل کرده و یک روش جامع را توسعه دهیم.

در این پژوهش برای رفع برخی از مشکلات روش‌های موجود، دو رویکرد محلی و سراسری را با هم ترکیب نموده‌ایم. برای ترکیب این دو رویکرد ابتدا داده‌ها را بر اساس برخی مولفه‌ها مانند فاصله و با استفاده از روش GOTHiC مدل کردیم. بر اساس مدل بدست آمده و مقدار اولیه Read counts فعل و انفعالات، شبکه عصبی اتوانکدر را آموزش می‌دهیم. در واقع در اینجا شبکه عصبی اتوانکدر مدلی را برای تخمین مقدار دقیق‌تر Read counts فعل و انفعالات بر اساس فعل و انفعالات نزدیک به هم (رویکرد محلی)، ایجاد می‌کند. در نهایت هم پس از تخمین مقدار Read counts تمام فعل و انفعالات، روش پیشنهادی فعل و انفعالات معنی‌دار را با استفاده از آزمون دوجمله‌ای شناسایی کرده است. برای ارزیابی روش ارائه شده از آنجایی که از روش‌های معمول ارزیابی نمی‌توان استفاده کرد، از دو مفهوم وابستگی بین مولفه‌ها و همچنین تعداد فعل و انفعالات معنی‌دار استفاده کردیم. لازم به ذکر است که تعداد فعل و انفعالات معنی‌دار به غیر از ارزیابی در تحلیل داده‌های Hi-C نیز کاربرد دارد.

در این پژوهش روش توسعه داده شده را با روش‌های GOTHiC و Fit-Hi-C مقایسه نموده‌ایم. همچنین برای ارزیابی از ضریب همبستگی پیرسون بین معیار فاصله و مقادیر تخمین زده شده و همچنین تعداد فعل و انفعالات معنی‌دار استفاده کردیم. بر اساس ضریب همبستگی پیرسون می‌توان

نتیجه گرفت که این روش به خوبی توانسته است داده‌های Hi-C را بر اساس فرض اولیه و اصلی مساله مدل کند. با وجود اینکه روش ارائه شده از شبکه عصبی عمیق استفاده می‌کند، با این حال این روش توانسته داده‌های Hi-C را براساس معیار فاصله به خوبی مدل کند و رابطه این دو را به خوبی نشان دهد. با ارزیابی تعداد فعل و انفعالات معنی‌دار به این نتیجه رسیدیم که روش ارائه شده عملکرد به مراتب بهتری را نسبت به روش GOTHiC و عملکرد مشابه‌ای را نسبت به روش Fit-Hi-C داشته است. بر اساس نتایج حاصل شده می‌توان نتیجه گرفت که روش پیشنهادی عملکرد قابل قبولی داشته و توانسته تعداد زیادی فعل و انفعال معنی‌دار شناسایی کند. به طور کلی بر اساس نتایج بدست آمده به این نتیجه رسیدیم که با استفاده از این روش ترکیبی (ترکیبی از مدل‌های آماری و شبکه عصبی) می‌توان به خوبی داده‌های Hi-C را مدل نموده و داده‌های Hi-C را نرمال سازی کرد.

لازم به ذکر است که در روند انجام پژوهش به این نتیجه رسیده‌ایم که داده‌ها Hi-C خود از پروتکل‌های مختلفی استخراج می‌شوند. از آنجایی که هر پروتکل دارای ساختار منحصر به فرد است، از این رو روش‌های توسعه داده شده بهتر است، توانایی مدل کردن این ساختارهای متفات را داشته باشند. با توجه به این موضوع ما این روش را با استفاده از شبکه عصبی اتوانکدر توسعه داده‌ایم. در واقع شبکه عصبی اتوانکدر دادگان بر اساس خود داده‌ها مدل می‌کند. لذا روش ارائه در این پژوهش به هیچ یک از انواع داده‌های Hi-C حساس نیست.

علاوه بر این، همراه با توسعه روش پیشنهادی، ابزاری به نام MHiC را برای حذف نویز و شناسایی فعل و انفعالات معنی‌دار در محیط R توسعه داده‌ایم. در این ابزار علاوه بر روش پیشنهادی، روش‌های دیگری نیز برای حذف نویز از داده‌های Hi-C پیاده سازی شده است. به غیر از حذف نویز این ابزار داده‌های Hi-C را با استفاده از Contact map diagram و Arc diagram نمایش می‌دهد. به طور کلی در ابزار ارائه شده مجموعه روش‌های موجود برای حذف نویز را در یک ابزار گردآوردی کرده‌ایم و همچنین امکان بصری‌سازی داده‌های Hi-C را برای پژوهشگران فراهم نموده‌ایم.

۵-۲ پیشنهادات و کارهای آینده

در فرایند انجام پژوهش با چالش‌های زیادی مواجه شدیم. بر اساس این چالش‌ها و ایده‌های مطرح شده، می‌توان راه‌کارهای زیر را مورد توجه قرار داد.

- یکی از چالش‌های اصلی موجود، بار محاسباتی بسیار زیاد این داده‌ها می‌باشد که برای این مساله می‌توان از پردازش موازی در طراحی الگوریتم مورد نظر استفاده نمود.
- چالش دیگری که با آن روبرو بوده‌ایم، انتخاب یکی از انواع شبکه عصبی و هماهنگ‌سازی داده‌های Hi-C با آن شبکه عصبی بود. لذا استفاده از دیگر شبکه عصبی‌ها در صورت هماهنگ سازی با داداه‌های Hi-C ممکن است.
- انتخاب روشی برای مدل‌سازی اولیه خود نیز یکی از چالش‌های مساله بوده است. لذا به دلیل اینکه روش‌های بسیار زیادی برای مدل کردن دادگان به صورت آماری وجود دارد، می‌توان مدل آماری استفاده شده در روش پیشنهادی را بهبود داد و یا آن را برای گرفتن نتیجه بهتر تغییر داد.
- برای درک بهتر از دادگان نیاز است که به ابزار توسعه داده شده بخش‌های مختلفی برای بصری سازی و همچنین تحلیل دادگان اضافه نمود. به عبارت دیگر بخش جدیدی به ابزار MHiC برای تحلیل دادگان نرمال شده اضافه شود.

فهرست واژگان

ا

آدنین.....Adenine

اتوانکدر.....Autoencoder

آزمون دوجمله‌ای .....Binomial test

پ

ثبت پیکربندی کروموزوم.....
Chromosome conformation capture

ت

تیامین.....Thymine

ج

جفت باز.....Base Pair

خ

خوشه‌بندی کی-میانگین.....K-means

د

دی ان ای.....Deoxyribonucleic acid

ر

رگرسیون پواسون.....Poisson regression

رگرسیون دو جمله‌ای منفی..... Negative binomial regression

س

سیتوزین.....Cytosine

ش

شبکه‌های عصبی بازگشتی..... Recurrent Neural Network

شبکه‌های عصبی عمیق..... Deep Neural Networks

ف

فعل و انفعالات .....Interactions

فعل و انفعالات معنی‌دار.....Significant Interactions

فعل و انفعالات بین کروموزومی.....Inter Interactions

فعل و انفعالات داخل کروموزومی...Intra Interactions

گ

گوانین.....Guanine

مراجع

[1] Y. Yuan *et al.*, "Cancer type prediction based on copy number aberration and chromatin 3D structure with convolutional neural networks," *BMC Genomics*, vol. 19, 2018.

[2] F. Jiang *et al.*, "Artificial intelligence in healthcare: Past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4. pp. 230–243, 2017.

[3] J. S. Baxter *et al.*, "Capture Hi-C identifies putative target genes at 33 breast cancer risk loci," *Nat. Commun.*, vol. 9, no. 1, 2018.

[4] G. Orlando *et al.*, "Promoter capture Hi-C-based identification of recurrent noncoding mutations in colorectal cancer," *Nature Genetics*, vol. 50, no. 10. pp. 1375–1380, 2018.

[5] N. Servant, N. Varoquaux, E. Heard, E. Barillot, and J. P. Vert, "Effective normalization for copy number variation in Hi-C data," *BMC Bioinformatics*, vol. 19, no. 1, 2018.

[6] R. Jia, P. Chai, H. Zhang, and X. Fan, "Novel insights into chromosomal conformations in cancer," *Molecular Cancer*, vol. 16, no. 1. 2017.

[7] H.-J. Wu and F. Michor, "A computational strategy to adjust for copy number in tumor Hi-C data," *Bioinformatics*, vol. 32, no. 24, pp. 3695–3701, 2016.

[8] T. Cremer and C. Cremer, "Rise, fall and resurrection of chromosome territories: A historical perspective. Part I. The rise of chromosome territories," *European Journal of Histochemistry*, vol. 50, no. 3. pp. 161–176, 2006.

[9] M. Vietri Rudan *et al.*, "Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture," *Cell Rep.*, vol. 10, no. 8, pp. 1297–1309, 2015.

[10] S. Sati and G. Cavalli, "Chromosome conformation capture technologies and their impact in understanding genome function," *Chromosoma*, vol. 126, no. 1. pp. 33–44, 2017.

[11] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, "Capturing chromosome conformation," *Science (80-.).*, vol. 295, no. 5558, pp. 1306–1311, 2002.

[12] J. Dekker, "The three 'C' s of chromosome conformation capture: Controls, controls, controls," *Nat. Methods*, vol. 3, no. 1, pp. 17–21, 2006.

[13] M. Simonis *et al.*, "Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)," *Nat. Genet.*, vol. 38, no. 11, pp. 1348–1354, 2006.

[14] J. Dostie *et al.*, "Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements," *Genome Res.*, vol. 16, no. 10, pp. 1299–1309, 2006.

[15] E. Lieberman-Aiden *et al.*, "Comprehensive mapping of long-range interactions

reveals folding principles of the human genome," *Science (80-.).*, vol. 326, no. 5950, pp. 289–293, 2009.

[16] G. Li *et al.*, "Chromatin interaction analysis with paired-end tag (ChIA-PET) sequencing technology and application," *BMC Genomics*, vol. 15, 2014.

[17] F. Ay, T. L. Bailey, and W. S. Noble, "Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts," *Genome Res.*, vol. 24, no. 6, pp. 999–1011, 2014.

[18] M. Hu, K. Deng, S. Selvaraj, Z. Qin, B. Ren, and J. S. Liu, "HiCNorm: Removing biases in Hi-C data via Poisson regression," *Bioinformatics*, vol. 28, no. 23, pp. 3131–3133, 2012.

[19] B. Mifsud *et al.*, "GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data," *PLoS One*, vol. 12, no. 4, 2017.

[20] S. S. P. Rao *et al.*, "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, vol. 159, no. 7, pp. 1665–1680, 2014.

[21] S. Wingett *et al.*, "HiCUP: pipeline for mapping and processing Hi-C data," *F1000Research*, 2015.

[22] J. Cairns *et al.*, "CHiCAGO: Robust detection of DNA looping interactions in Capture Hi-C data," *Genome Biol.*, vol. 17, no. 1, 2016.

[23] M. Forcato, C. Nicoletti, K. Pal, C. M. Livi, F. Ferrari, and S. Bicciato, "Comparison of computational methods for Hi-C data analysis," *Nat. Methods*, vol. 14, no. 7, pp. 679–685, 2017.

[24] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2.," *Nat. Methods*, vol. 9, no. 4, pp. 357–9, 2012.

[25] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, no. 3, 2009.

[26] M. W. Schmid, S. Grob, and U. Grossniklaus, "HiCdat: A fast and easy-to-use Hi-C data analysis tool," *BMC Bioinformatics*, vol. 16, no. 1, 2015.

[27] N. Servant *et al.*, "HiC-Pro: An optimized and flexible pipeline for Hi-C data processing," *Genome Biol.*, vol. 16, no. 1, 2015.

[28] S. Heinz *et al.*, "Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities," *Mol. Cell*, vol. 38, no. 4, pp. 576–589, 2010.

[29] S. V. Ulianov, A. A. Gavrilov, and S. V. Razin, "Nuclear Compartments, Genome Folding, and Enhancer-Promoter Communication," *Int. Rev. Cell Mol. Biol.*, vol. 315, pp. 183–244, 2015.

[30] A. T. L. Lun and G. K. Smyth, "diffHic: A Bioconductor package to detect differential genomic interactions in Hi-C data," *BMC Bioinformatics*, vol. 16, no. 1, 2015.

[31] J. M. F. Chebli, *Effect of azathioprine or mesalazine therapy on incidence of re-hospitalization in sub-occlusive ileocecal Crohn's disease patients*, vol. 19. 2013.

[32] MTW and J. S. Long, “Regression Models for Categorical and Limited Dependent Variables.,” *J. Am. Stat. Assoc.*, vol. 92, no. 440, p. 1655, 1997.

[33] G. Rushin, C. Stancil, M. Sun, S. Adams, and P. Beling, “Horse race analysis in credit card fraud - Deep learning, logistic regression, and Gradient Boosted Tree,” in *2017 Systems and Information Engineering Design Symposium, SIEDS 2017*, 2017, pp. 117–121.

[34] D. Kwon, H. Kim, J. Kim, S. C. Suh, I. Kim, and K. J. Kim, “A survey of deep learning-based network anomaly detection,” *Cluster Computing*, pp. 1–13, 2017.

[35] E. L. Paula, M. Ladeira, R. N. Carvalho, and T. Marzagão, “Deep learning anomaly detection as suppor fraud investigation in Brazilian exports and anti-money laundering,” in *Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016*, 2017, pp. 954–960.

[36] R. Xie, J. Wen, A. Quitadamo, J. Cheng, and X. Shi, “A deep auto-encoder model for gene expression prediction,” *BMC Genomics*, vol. 18, 2017.

[37] T. Schulz, J. Stoye, and D. Doerr, “GraphTeams: A method for discovering spatial gene clusters in Hi-C sequencing data,” *BMC Genomics*, vol. 19, 2018.

[38] I. Irastorza-Azcarate, R. D. Acemel, J. J. Tena, I. Maeso, J. L. Gómez-Skarmeta, and D. P. Devos, “4Cin: A computational pipeline for 3D genome modeling and virtual Hi-C analyses from 4C data,” *PLoS Comput. Biol.*, vol. 14, no. 3, 2018.

[39] C. J. Cameron, J. Dostie, and M. Blanchette, “Estimating DNA-DNA interaction frequency from Hi-C data at restriction-fragment resolution,” *bioRxiv*, p. 377523, 2018.

[40] J. R. Dixon *et al.*, “Topological domains in mammalian genomes identified by analysis of chromatin interactions,” *Nature*, vol. 485, no. 7398, pp. 376–380, 2012.

[41] M. Hu *et al.*, “Bayesian Inference of Spatial Organizations of Chromosomes,” *PLoS Comput. Biol.*, vol. 9, no. 1, 2013.

Abstract

Today, medical information analysis can raise our understanding of the structure of the human body and its influential factors. Hence, researchers in the field of medicine are trying to identify these factors to prevent the spread of diseases. One of these influential factors is the structure and method of placing DNA strands in the three-dimensional space of chromosomes. The main idea behind this topic proposed where, in a complex DNA strand, interactions between two DNA regions that have a huge impact on body function when they are close to the spatial location. Therefore, various laboratory protocols have developed to obtain the structural information of the chromosomes. One of these protocols is Hi-C. This laboratory protocol has many systematic and laboratory errors. As a result, to use this information and to better understand body function, we need to remove noise and extract meaningful information from these data is essential. So far, many studies have been done in this regard, and various statistical methods have been developed to solve this problem. Generally, these methods have used a global statistical approach to detect noise. However, the process of noise recognition can be improved by considering a local approach. Therefore, in this abstract, merged local and global approaches together. For this purpose, proposed a method based on a deep neural network that called Auto encoder, due to the ability of this network to eliminate noise. The proposed method initially models Hi-C data statistically. Then, this method improves the statistical model with respect to the impact of data on each other by using the neural network, in other words, generates a new model using the neural network.

The simulation results show that the proposed method has had better performance with 3,771 credible interactions than the reference method (GOTHiC) with 3,772 valid interactions in the proposed method. Also, the correlation coefficient between the distance component and the number of interactions between the two regions in the proposed method and the reference method was respectively -0.1951 and -0.3100, respectively. As a result of the proposed method, the dependence of the number of interactions between the two area is more adherent to the distance between two areas. In general, in this study, we have shown that it is possible to simulate Hi-C data using the

neural network and, based on the model created by the neural network, eliminates the noise, as well as the verb Meaningful influence has been identified in Hi-C data.
In addition, along with the development of the proposed method, we developed a tool called MHiC to eliminate noise and identify meaningful interactions in the R environment. In addition to the proposed method, we have implemented GOTHiC, HiCNorm and FitHiC methods to eliminate noise and detect meaningful interactions. Unlike existing implementations, each method implemented in this tool is capable of receiving input from HiC-Pro, HOMER, HiCUP sources as well as inputs designed for the HiCNorm method. The tool also displays Hi-C data using the Contact map diagram and the Arc diagram. The provided tool in this data enables the visualization of Hi-C data and provides a set of available techniques for noise elimination.

Keywords: Hi-C, Deep Neural Network, noise removal, Bioinformatics, regression

Shahrood University of Technology

Faculty of Computer Engineering

M.Sc. Thesis in Artificial Intelligence

A model for identifying and eliminating noise and bias from Hi-C data

By: Saman Khakmardan

Supervisor:
Dr. Mohsen Rezvani and Dr. Ali Akbar Pouyan

Advisor:
Dr. Mansoor Fateh and Dr. Hamid Alinejad Rokny

February 2019