مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره پنجم شماره دوم. ص۳۱۳-۳۰۴، ۱۳۹۷

کاوش و استخراج برهم‌کنش‌های پروموتر/انهنسر بالقوه در ژنوم افراد سرطانی با استفاده از یک الگوریتم چندهدفه تکاملی

محمدجوادحسین پور[۱]، حمید پروین[۲،۶]، صمد نجاتیان[۳،۵*]، وحیده رضایی[۴،۵]

•دریافت مقاله: ۹۶/۷/۳۰ •پذیرش مقاله: ۹۷/۳/۲۲

مقدمه: سرطان به عنوان یکی از شایع‌ترین انواع بیماری‌ها، سلامت بسیاری از انسان‌ها را تحت تأثیر قرار داده است. هدف اصلی در این مقاله، ارائه یک الگوریتم تکاملی چندهدفه می‌باشد. این الگوریتم، با استفاده از اطلاعات برهم‌کنش بین ژنومی در کروموزوم‌های افراد مبتلا به سرطان، ناحیه‌های بالقوه پروموتر/انهنسر را کشف و استخراج می‌کند. استخراج صحیح این ناحیه‌ها می‌تواند به علم پزشکی در تشخیص زودهنگام بیماری سرطان کمک کند.

روش: پژوهش حاضر از نوع کاربردی و توصیفی می‌باشد. در این پژوهش از مجموعه داده Hi-C شامل اطلاعات مربوط به برهم‌کنش‌های بین ژنومی، در سلول GM12878 استفاده شد. جهت کشف و استخراج پروموتر/انهنسرهای بالقوه از الگوریتمی تکاملی و چندهدفه استفاده شد. الگوریتم مذکور با استفاده از نرم‌افزار متلب پیاده‌سازی گردید. همچنین کارایی این الگوریتم نیز، با استفاده از دو معیار مورد ارزیابی قرار گرفت. معیار اول، تابع تناسبی است که میزان برهم‌کنش‌های بین ژنومی نسبت به طول نواحی ژنوم را محاسبه می‌کند و معیار دوم تعداد پروموتر/انهنسرهای بالقوه کشف شده می‌باشد.

نتایج: نتایج و مقایسه‌های انجام گرفته در این پژوهش، نشان از کارایی بالا و بهینه بودن روش پیشنهادی در کشف پروموتر/انهنسر با طول متغیر نسبت به روش HiC-Pro می‌باشد؛ بنابراین روش پیشنهادی می‌تواند پروموتر/انهنسرهای بالقوه‌ای را کشف کند که روش HiC-Pro قادر به کشف آن نیست.

نتیجه‌گیری: با توجه به نتایج به دست آمده، الگوریتم پیشنهادی قادر به کشف و استخراج بهینه پروموتر/انهنسرهای بالقوه با طول متغیر می‌باشد، که می‌تواند در تشخیص زودهنگام سرطان کمک شایانی به علم پزشکی کند.

کلید واژه‌ها: پروموتر، انهسر، مجموعه داده Hi-C، الگوریتم شبیه‌سازی تبرید چند هدفه MOSA، روش HiC-Pro

•**ارجاع:** حسین پورمحمدجواد، پروین حمید، نجاتیان صمد، رضایی وحیده کاوش و استخراج برهم‌کنش‌های پروموتر/انهنسر بالقوه در ژنوم افراد سرطانی با استفاده از یک الگوریتم چندهدفه تکاملی. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۷؛ ۵(۲): ۳۱۳-۳۰۴.

۱. دانشجوی دکتری دانشکده مهندسی کامپیوتر، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران
۲. دکتری مهندسی کامپیوتر- هوش مصنوعی استادیار، دانشکده مهندسی کامپیوتر، واحد نورآباد ممسنی، دانشگاه آزاد اسلامی، نورآباد ممسنی، ایران
۳. دکتری مهندسی برق، استادیار، دانشکده مهندسی برق، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران
۴. دکتری ریاضیات، استادیار، دانشکده ریاضیات، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران
۵. عضو باشگاه پژوهشگران جوان و نخبگان، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران
۶. عضو باشگاه پژوهشگران جوان و نخبگان، واحد نورآباد ممسنی، دانشگاه آزاد اسلامی، نورآباد ممسنی، ایران

*** نویسنده مسئول:** دانشکده مهندسی برق، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران.•

شماره تماس: ۰۹۱۷۳۴۱۴۰۰۸

• **Email:** nejatian@iauyasooj.ac.ir

مقدمه

امروزه در دانش پزشکی میزان داده‌های مربوط به علائم بیماران سرطانی و روش‌های کمکی برای تشخیص این بیماری‌ها، بسیار وسیع و گسترده شده است، به طوری که معمولاً تحلیل و در نظرگرفتن کلیه عوامل دخیل، دشوار به نظر می‌رسد. در واقع سرطان یکی از شایع‌ترین انواع بیماری‌ها در انسان است که خیلی از انسان‌ها را در مراحل مختلف زندگی تحت تأثیر قرار می‌دهد. در سال‌های اخیر، میزان شیوع سرطان روند رو به رشدی داشته و بررسی داده‌ها نشان می‌دهد که میزان بقای بیماران تا پنج سال پس از تشخیص، ۸۸ درصد و ده سال پس از تشخیص ۸۰ درصد بوده است [۱]؛ بنابراین لازم است که این بیماری سریع تشخیص داده شود. یکی از روش‌های تشخیص زودهنگام سرطان کشف برهم‌کنش‌های غیر معمول موجود در ژنوم افراد بیمار و استخراج نواحی درگیر در این برهم‌کنش‌ها است. در واقع ژنوم‌های انسانی دارای دو عنصر پروموتر (Promoter) و انهنسر (Enhancer) هستند. حال، زمانی‌که یک ژن بخواهد بیان (Gene Expression) شود و فعالیتی را انجام دهد، این دو عنصر باید با یکدیگر برهم‌کنش (Interaction) داشته باشند. دراین‌صورت اگر آن ژن، مربوط به تومور سرطانی در شخص بیمار باشد، باعث می‌شود که آن ژن در آن فرد سرطانی بیشتر فعالیت کرده و گسترش پیدا کند [۲]. حال تشخیص دادن این نواحی راه‌انداز/ انهنسر مسئله مهمی است که به علم پزشکی در تشخیص زودهنگام سرطان کمک کرده و همچنین در درمان این بیماری می‌تواند مؤثر باشد. در این راستا، تکنولوژیHi-C به جهت کشف برهم‌کنش‌های بیشتر در ژنوم به وجود آمد. این تکنولوژی برای اولین بار توسط آیدین و همکاران جهت شبیه‌سازی کردن همه برهم‌کنش‌های کروموزومی در یک آزمایش خاص مطرح گردید. در حقیقت تکنولوژی Hi-C به همراه ابزارهای آن برهم‌کنش‌های بین نقاط مختلف ژنوم را در یک ساختار سه بعدی ارائه می‌دهد [۳]. بدین معنی که از نظر ساختاری چه ناحیه‌ای از ژنوم با چه ناحیه‌ای از ژنوم دیگر برهم‌کنش دارد. این اطلاعات می‌توانند دانش زیادی را در زمینه ساختار برهم‌کنش ژنوم ارائه کند و بسیار گران قیمت نیز می‌باشند. از جمله ابزارهای Hi-C می‌توان به (Humer, HiCup, HiC-Pro, etc..) اشاره کرد، که در تحقیقات گذشته [۴–۷] به‌منظور کشف پروموتر/انهنسر و برهم‌کنش‌های موجود بین آن‌ها مورد استفاده قرار گرفته است، که در بین آن‌ها کارایی ابزار HiC-Pro از دیگر ابزارها بالاتر می‌باشد.

در سال‌های گذشته پژوهش‌های زیادی در این زمینه صورت گرفته است. Jin و همکاران، در مقاله‌ای نگاشتی از برهم‌کنش‌های سه بعدی کروماتین‌ها در سلول‌های انسانی ارائه کردند. در این پژوهش آن‌ها یک نگاشت از برهم-کنش‌های کروماتین ایجاد شده در فیبروبلاست‌های انسانی با استفاده از یک روش تحلیل C۳ با ژنوم گسترده (Hi-C) به دست آوردند. همچنین بیشتر از یک میلیون برهم‌کنش‌های کروماتین با دامنه طویل در فرگمنت‌های 10-15kb تعیین شد و اصول کلی سازمان کروماتین را در انواع متفاوت ویژگی‌های ژنومی مشخص گردید. مشاهدات آن‌ها پیشنهاد می‌دهد که چشم انداز کروماتین سه بعدی (زمانی که در یک نوع سلولی خاص ایجاد شد) تقریباً ثابت است و می‌تواند بر انتخاب یا فعال‌سازی ژن‌های مقصد توسط یک فعال ساز رونویسی موجود در همه جا به یک روش خاص سلول تأثیرگذار باشد [۸]. Cabreros و همکاران، الگوریتم تشخیص اجماعی بر روی داده‌های Hi-C جهت کشف اجماعی از نواحی در ساختار سه بعدی دی‌ان‌ای موش و انسان به کار بردند. الگوریتم پیشنهادی آن‌ها قادر به کشف تعداد متغیری از اجماع بود. همچنین این الگوریتم، می‌توانست اجماعی از مکان‌های دی‌ان‌ای که در دور و نزدیک از هم قرار دارند، را به صورت دنباله‌ای کشف کند [۹]. Mifsud و همکاران، در مقاله‌ای نگاشت تماس‌های پروموتر با طول دامنه بالا در سلول‌های انسانی با ضبط Hi-C با کیفیت بالا ارائه کردند. آن‌ها در این پژوهش در آزمایشگاه ژنتیک از ضبط Hi-C (Chi-C) (یک آزمایش ژنوم اتخاذ شده) استفاده کردند تا تعاملاتی با دامنه طویل، تقریباً ۲۲۰۰۰ پروموتر را در ۲ نوع سلول‌های خون انسان مورد سنجش قرار دهند. این رویکرد می‌تواند توالی‌سازی عمیق ناحیه خاص را مشخص کرده و باعث شود تا نگاشت برهم‌کنش‌ها با کیفیت بالا بین مناطق انجام شود [۱۰]. Servant و همکاران، در مقاله‌ای ابزاری به نام HiC-Pro جهت کشف برهم‌کنش‌های بین ژنومی در کروموزوهای انسانی ارائه دادند. ابزار پیشنهادی آن‌ها می‌توانست برهم‌کنش‌های یک ژنوم خاص با دیگر ژنوم را به دست آورد. همچنین HiC-Pro قابلیت اجرا به صورت موازی نیز داشت، که منجر به سرعت بالای آن در آنالیز مجموعه داده‌های Hi-C شده بود [۵]. Fotuhi و همکاران، در مقاله‌ای یک الگوریتم خوشه‌بندی چند منظوره برای مجموعه داده‌های به دست آمده از پیکربندی کروموزوم جهت شناسایی مدل‌های برهم‌کنش‌های کروموزومی

ارائه کردند. در این مقاله روش Arboretum-Hi-C (روش خوشه‌بندی طیفی چند منظوره) برای شناسایی جنبه‌های مشترک و در زمینه خاصی از ساختار ژنوم ارائه گردید. در مقایسه با روش خوشه‌بندی استاندارد، روش Arboretum-Hi-C الگوهای بیولوژیکی را که برای نگهداری سازگارتر بودند، به دست آورد. همچنین خوشه‌هایی که این الگوریتم خوشه‌بندی ارائه داد بهتر از روش‌های استاندارد بود [۱۱].

Charalamposو همکاران، ابزار HiC-bench جهت آنالیز مجموعه داده‌های Hi-C ارائه کردند. این ابزار، علاوه بر کشف پروموتر/انهنسر، برهم‌کنش‌های بین آن‌ها را نیز استخراج می‌کند. همچنین ابزار پیشنهادی قادر به اجرا به صورت موازی بر روی چند سیستم، مصورسازی نتایج (visualize)،کشف برهم‌کنش‌های دورن و بین کروموزومی و اجرا در سریع‌ترین زمان ممکن می‌باشد [۱۲].

اما مشکل اصلی موجود در مقالات مورد بررسی این است‌که، فناوری Hi-C و ابزارهای آن تنها می‌توانند فرگمنت‌های با طول ثابت(Fix-bin size) از ژنوم جهت تشکیل پروموتر و انهنسر کشف و استخراج کرده و برهم‌کنش‌های بین آن‌ها را شمارش کنند. در این صورت خروجی این ابزار و روش‌های ارائه شده قبلی، که با فرگمنت‌های با طول ثابت کار می‌کنند، نمی‌توانند بهینه باشد. در این صورت ارائه یک راه‌حل که بتواند پروموتر/انهنسر با طول متغیر را کشف و استخراج کرده و برهم‌کنش‌های بین آن‌ها را شمارش کنند، حس می‌شود.

در این تحقیق یک الگوریتم چندهدفه مبتنی بر شبیه‌سازی تبرید ارائه شد که بتواند فرگمنت‌های با طول متغیر از ژنوم که پروموتر و انهنسر بالقوه(potential promoter/enhancer) می‌باشند را کشف و استخراج کرده و برهم‌کنش‌های بین آن‌ها را شمارش کند.

روش

پژوهش حاضر از نوع کاربردی و توصیفی می‌باشد. در این پژوهش، با استفاده از الگوریتمی چندهدفه مبتنی بر شبیه‌سازی تبرید نواحی با طول متغیر پروموتر/انهنسر را ژنوم بیماران سرطانی کشف و استخراج شد و تمامی عملیات الگوریتم پیشنهادی بر روی مجموعه داده Hi-C، که اطلاعات مربوط به برهم‌کنش‌های بین ژنومی در سلول GM12878 انسانی می‌باشد، انجام گرفت.

طرح مسئله

اگر مجموعه داده را به صورت n قلم داده در نظر گرفته شود. $D = \{g_1 \to g_{1 \le i \le n}, g_2 \to g_{1 \le i \le n}, \dots, g_n \to g_{1 \le i \le n}\}$ در اینجا g نشان‌دهنده یک ناحیه از ژنوم موجود در یک کروموزوم است. حال هدف مسئله یافتن اقلام پرتکرای مانند $P \to E$ که P و E به شکل زیر باشد:

$$P = \{g_{i_1}, g_{i_1+1}, \dots, g_{i_2}\}, i_1 < i_2$$
$$E = \{g_{i_3}, g_{i_3+1}, \dots, g_{i_4}\}, i_3 < i_4$$

در اینجا g_{i_1} تا g_{i_2} نواحی بهم پیوسته از ژنوم یک کروکوزوم می‌باشد که مشخص کننده یک پروموتر بالقوه و g_{i_3} تا g_{i_4} نیز نواحی بهم پیوسته از ژنوم یک کروکوزوم است که مشخص کننده یک انهنسر بالقوه می‌باشد. مجموعه داده مورد استفاده در این تحقیق مجموعه داده Hi-C می‌باشد، که این مجموعه داده برهم‌کنش‌های بین نقاط مختلف ژنوم را در افراد سرطانی نشان می‌دهد. اگر مجموعه داده Hi-C را برابر با D در نظر گرفته شود. مسئله مورد نظر یافتن رابطه برهم‌کنش $P \to E$ می باشد، که هر یک از نواحی P با تمامی نواحی E برهم-کنش داشته باشند. که در این رابطه P پروموتر و E انهنسر در نظر گرفته می‌شود.

کشف پروموتر/انهنسر چندهدفه

در این بخش روش پیشنهادی به منظور کاوش و استخراج پروموتر/انهنسر در ژنوم افراد سرطانی توضیح داده می‌شود. رویکرد پیشنهادی بر اساس الگوریتم شبیه‌سازی تبرید چندهدفه می‌باشد، که در این بخش هر یک از اهداف مورد استفاده در الگوریتم پیشنهاد شده آورده شد. به‌منظور تشریح هریک از اهداف فرض می‌شود که $P \to E$ مطابق با آنچه در بخش قبل توضیح داده شد، یک جواب مسئله می‌باشد.

هدف اول: ضریب نسبیت(Ratio)

ضریب نسبیت، برابر با جمع تمامی برهم‌کنش‌های هر قلم داده در مجموعه P با هر قلم داده در مجموعه E تقسیم بر جمع تمامی برهم‌کنش‌های موجود در مجموعه P با دیگر قلم داده های موجود در مجموعه D می‌باشد. مقدار ضریب اطمینان در بازه [۰،۱] می‌باشد و هر چقدر این مقدار بیشتر باشد ضریب نسبیت بهتری را ارائه می‌دهد.

$$Ratio\ (P \to E) = \frac{\sum_{m=i_1}^{i_2} \sum_{p=i_3}^{i_4} (g_m \to g_p)}{\sum_{m=i_1}^{i_2} \sum_{o=1}^{n} (g_m \to g_o)} \qquad (۱)$$

هدف دوم: ضریب پیوستگی

ضریب پیوستگی، برابر با حاصل‌ضرب طول تمامی قلم داده‌های موجود در مجموعه موردنظر (P یا E) در محدودیت پیوستگی (Continuous Limitation)CL آن مجموعه.

$$Sequence(P) = Length(P) \times CL(P) \qquad (۲)$$

محدویت پیوستگی اقلام داده یک متغیر نوع منطقی (Boolean) می‌باشد، که در صورت پیوسته بودن اقلام مجموعه مورد نظر مقدار ۱ و در غیر این‌صورت مقدار ۱- را می‌گیرد. پیوسته بودن، بدین معنی است که مجموعه‌ای از ژنوم که تشکیل پروموتر یا انهسر را داده‌اند به صورت ناحیه‌ای پیوسته از کروموزوم مربوطه باشند.

تابع تناسب(Fitness)

تابع تناسب برابر است با حاصل تقسیم ضریب نسبیت بر مجموع ضریب پیوستگی مجموعه‌های P و E. در حقیقت این معیار ضریب برهم‌کنش‌های بین مجموعه اقلام P با مجموعه اقلام E را محاسبه می‌کند. مقدار تابع تناسب عددی بین ۱- تا ۱ است که هر چه به ۱ نزدیک‌تر باشد مقدار تابع تناسب بهتر خواهد بود و نشان‌دهنده وجود برهم‌کنش بیشتر بین هر دو نواحی P و E می‌باشد.

$$Fitness(P \to E) = \left(\frac{Ratio(P \to E)}{(Sequence(P) + Sequence(E))}\right) \qquad (۳)$$

الگوریتم پیشنهادی

تعداد جمعیت اولیه در الگوریتم پیشنهادی برابر با ۲۵، بیشینه تکرار برابر با ۱۰۰، تعداد زیر تکرار برابر با ۲۰ و تعداد همسایگان قابل کشف در نزدیکی جواب بهینه ۱۰ می‌باشد. ورودی الگوریتم، مجموعه داده Hi-C حاصل از برهم‌کنش موجود بین ژنوم و تعداد کل فرگمنت‌های مورد استفاده در مجموعه داده می‌باشد.

در مرحله پیش پردازش ساده‌سازی مجموعه داده Hi-C انجام می‌گیرد. ساده‌سازی به‌منظور حذف تمامی برهم‌کنش‌های بین کروموزمی اجرا می‌شود. در این‌صورت مجموعه داده موردنظر شامل تمامی برهم‌کنش‌های درون کروموزمی می‌باشد. بعد از مرحله پیش‌پردازش، جمعیت اولیه به صورت تصادفی ساخته می‌شود، سپس این جمعیت توسط تابع تناسب مطابق با فرمول ۳ ارزیابی می‌شوند. پس از ارزیابی، فردی از جمعیت ورودی که نشان‌دهنده پروموتر/انهنسر با بهترین مقدار تابع تناسب است به‌عنوان بهترین پاسخ فعلی در نظر گرفته می‌شود. سپس دمای اولیه T_0 نیز تنظیم می‌شود که در این جاT_0= ۱۰ و میزان کاهش دما نیز $AlphA$= ۰/۹۹ در نظر گرفته شد. حال به اندازه بیشینه تکرار، مراحل زیر را تکرار شد. در نزدیکی جواب بهینه حاصله موجود به اندازه ۱۰ همسایه، جمعیت جدید را به وجود می‌آید، سپس توسط مقدار تابع تناسب هر یک از جمعیت تولیدی جدید را با بهترین جواب حاصله مقایسه شد. اگر پاسخ جدید از بهترین پاسخ یافته شده بهتر بود، پاسخ جدید را به عنوان بهترین پاسخ جدید در نظر گرفته شد؛ اما اگر بهتر نبود آن را با درصد احتمالی به صورت پاسخ مشروط در نظر گرفته می‌شود. سپس بهترین پاسخ یافته شده را به‌روزرسانی کرده و کاهش دما نیز به عنوان آخرین مرحله انجام می‌گیرد. حال این روند به اندازه تعداد تکرار ادامه پیدا می‌کند و در هر تکرار پاسخ بهینه ذخیره می‌شوند. در هر تکرار پاسخ بهینه، بهترین پروموتر/انهنسر را نشان می‌دهد، که می‌تواند مؤثر در بر هم‌کنش‌های بین ژنومی در بیماران سرطانی باشند.

ساختار یک عنصر از جمعیت ورودی الگوریتم پیشنهادی

ساختار هر فرد (Individual) از جمعیت ورودی الگوریتم پیشنهادی به صورت زیر می‌باشد. در ادامه هریک از قسمت‌های این ساختار توضیح داده می‌شود.

طول انهنسر	انهنسر	طول پروموتر	پروموتر

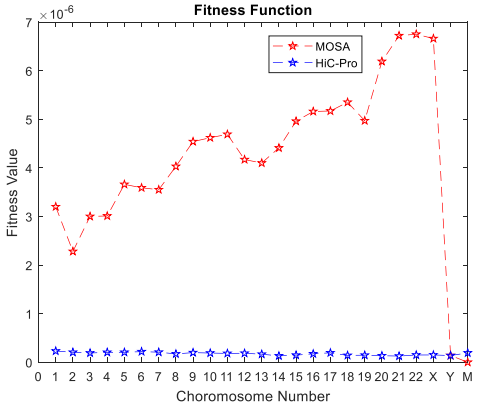
- راه انداز: مشخص‌کننده ابتدای ژنوم تشکیل دهنده پروموتر می‌باشد.
- طول راه‌انداز: مشخص کننده طول پروموتر می‌باشد، که به‌صورت تصادفی درنظر گرفته می‌شود.
- انهنسر: مشخص کننده ابتدای ژنوم تشکیل دهنده انهنسر می‌باشد.
- طول انهنسر: مشخص کننده طول انهنسر می‌باشد، که به‌صورت تصادفی در نظر گفته می‌شود.

مجموعه داده

در این تحقیق از مجموعه داده Hi-C استفاده شد. مجموعه داده مورد بررسی، حاوی ۲۵۰ میلیون رکورد مربوط به اطلاعات برهم‌کنش‌های بین ژنومی در سلول GM12878 افراد مبتلا به سرطان می‌باشد. سلول GM12878 یک سلول لیمفوبلاستوئیدی (lymphoblastoid) تولید شده از خون انسان با استفاده از تبدیل EBD(Epstein-Barr virus) می‌باشد. همچنین مجموعه داده مورد نظر در مطالعات متنوع ژنتیکی مانند [۱۰،۱۳] استفاده شده است.

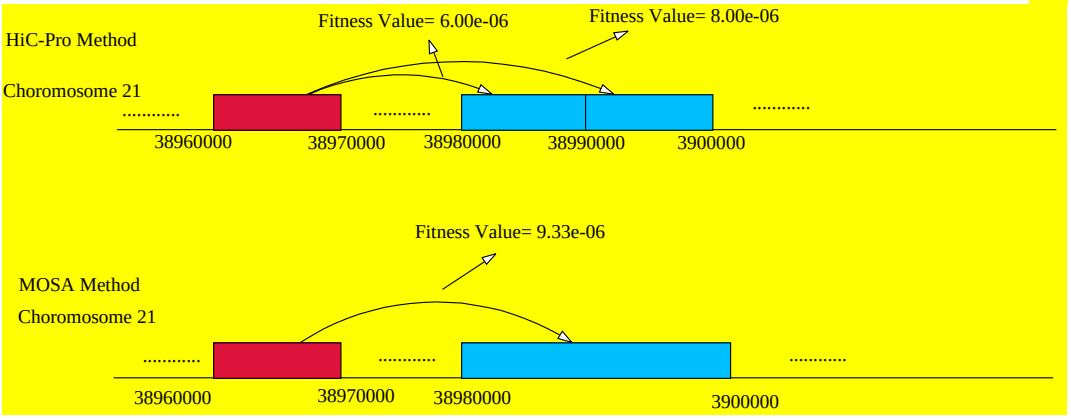
شکل ۱، میانگین تابع تناسب مربوط به روش پیشنهادی و روش HiC-Pro در همه کروموزوم‌ها بر روی مجموعه داده مورد نظر به‌تصویر کشیده شده است. مطابق این شکل روش پیشنهادی در همه کروموزوم‌ها بهتر از روش HiC-Pro می‌باشد. در ادامه، مطابق شکل ۲، روند محاسبه تابع ارزیابی با یک مثال در هر دو روش پیشنهادی و HiC-Pro توضیح داده شده است.

همانطور که شکل ۲(الف) نشان می‌دهد، ناحیه ۳۸۹۶۰۰۰۰ تا ۳۸۹۷۰۰۰۰ از کروموزوم ۲۱ را به‌عنوان یک پروموتر بالقوه تشخیص می‌دهد؛ اما ناحیه‌ای که به‌عنوان انهنسر در نظر گرفته است، ناحیه ۳۸۹۸۰۰۰۰ تا ۳۸۹۰۰۰۰۰ از کروموزوم ۲۱ می‌باشد. شکل ۲(ب)، یک فرد از جمعیت ورودی الگوریتم پیشنهادی را نشان می‌دهد. دراینجا مقدار اول ابتدای راه‌انداز، مقدار دوم طول راه‌انداز، مقدار سوم ابتدای انهنسر و مقدار چهارم طول انهنسر را بیان می‌کند. مطابق این ورودی، الگوریتم پیشنهادی، وجود برهم‌کنش بین پروموتر و انهنسر را بر روی مجموعه داده Hi-C بررسی کرده و مطابق آن تابع تناسب مربوط به آن را محاسبه می‌کند (در شکل ۲(ج) روند محاسبه تابع تناسب به تصویر کشیده شده است). حال اگر مقدار تابع تناسب مربوط به برهم‌کنش بین نواحی پروموتر/انهنسر جدید از مقادیر تابع تناسب هرکدام از برهم‌کنش‌های قبلی موجود در مجموعه داده بهتر بود، ناحیه جدید که حاصل ادغام دو ناحیه قبلی است را به عنوان انهنسر جدید در نظر می‌گیرد و برهم‌کنش بین این دو ناحیه را کشف و استخراج می‌کند. این نشان می‌دهد که، روش پیشنهادی در کشف نواحی پروموتر/انهنسر بالقوه توانایی بهتری نسبت به روش HiC-Pro دارد. چون روش HiC-Pro، ناحیه۳۸۹۶۰۰۰۰ تا ۳۸۹۷۰۰۰۰ را به عنوان پروموتر و ناحیه‌های ۳۸۹۸۰۰۰۰ تا ۳۸۹۹۰۰۰۰ و ۳۸۹۹۰۰۰۰ تا ۳۹۰۰۰۰۰ را به عنوان دو انهنسر که هریک با پروموتر دارای برهم‌کنش جداگانه هستند در نظر گرفته است و مقدار تابع تناسب حاصل از هر برهم‌کنش، به ترتیب 6.00E-06 و 8.00E-06 می‌باشد؛ اما روش پیشنهادی، ناحیه ۳۸۹۶۰۰۰۰ تا ۳۸۹۷۰۰۰۰ را به عنوان پروموتر و ناحیه ۳۸۹۸۰۰۰۰ تا ۳۹۰۰۰۰۰ را به عنوان انهنسر در نظر می‌گیرد، که مقدار تابع تناسب حاصل از برهم‌کنش بین آن‌ها 9.33E-06 می‌باشد و از هریک از دو مقدار تناسب ثبت شده در روش HiC-Pro بهتر است و این نشان می‌دهد که ناحیه‌های مذکور برهم‌کنش شدیدتری نسبت به ناحیه‌های کشف شده در روش HiC-Pro دارند؛ بنابراین ناحیه پروموتر/انهنسر کشف شده در روش پیشنهادی دقیق‌تر از نواحی کشف شده در روش HiC-Pro است.



شکل ۱: مقایسه روش پیشنهادی با روش HiC-Pro بر روی مجموعه داده

الف)



ب)

Individual Population in proposed method

36960000	10000	38980000	20000

ج)

HiC-Pro Method:
P: Promoter, E=Enhancer

$1)P(38960000-38970000)\rightarrow E(38980000-38990000)$

$2)P(38960000-38970000)\rightarrow E(38990000-3900000)$

$Fitness(1)=6.00\,e-06$

$Fitness(2)=8.00\,e-06$

MOSA Method:

$3)P(38960000-38970000)\rightarrow E(38980000-3900000)$

$$Ratio(P\rightarrow E)=\frac{(3+4)}{25}=0.28$$

$$CL(P)=1\,,CL(E)=1,Sequence\,(P)=1*10000=10000,Sequence(E)=1*20000=20000$$

$$Fitness(3)=\frac{(0.28)}{(10000+20000)}=9.33\,e-06$$

$$Fitness(3)>Fitness(1),Fitness(2)$$

شکل ۲: الف) یک نمونه کشف شده از پروموتر/انهنسر بالقوه با طول متغیر توسط روش پیشنهادی در کروموزوم ۲۱ سلول GM12878 در مقایسه با روش HiC-Pro که با فرگمنت‌های با طول ثابت کار می‌کند. ب) یک نمونه جمعیت ورودی الگوریتم جهت کشف پروموتر/انهنسر بالقوه شکل الف. ج) محاسبه تابع ارزیابی مربوط به پروموتر/انهنسر بالقوه کشف شده در روش پیشنهادی و روش HiC-Pro.

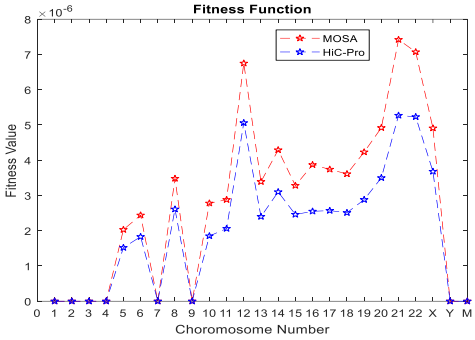
در ادامه روش پیشنهادی با روش HiC-Pro، مورد استفاده در [۴-۷]، مقایسه شده است. روش HiC-Pro از سایر روش‌ها، مانند HiCUP، HiCdat، Homer، HiClib و... در آنالیز مجموعه داده HiC قدرتمندتر می‌باشد. جدول ۱، تمامی برهم‌کنش‌های کشف شده توسط روش پیشنهادی که روش HiC-Pro قادر به کشف آن‌ها نبوده است را به تفکیک کروموزم نشان می‌دهد. همچنین میانگین مقدار تابع تناسب مربوط به هر کروموزوم حاصل از برهم‌کنش‌های کشف شده در الگوریتم پیشنهادی نسبت به میانگین مقدار تابع تناسب در روش HiC-Pro در همان کروموزوم نیز بیان شد.

مطابق با جدول ۱، روش پیشنهادی تعداد ۱۵۱ برهم‌کنش که روش HiC-Pro قادر به کشف آن‌ها نبوده را کشف کرده است. برهم‌کنش‌های کشف شده حاصل از وجود برهم-کنش‌های بین نواحی پروموتر/انهنسر بالقوه با طول متغیر می‌باشد که روش HiC-Pro قادر به کشف آن‌ها نیست.

شماره کروموزوم	تعداد برهمکنش روش پیشنهادی	میانگین مقدار تابع تناسب روش پیشنهادی	میانگین مقدار تابع تناسب روش HiC-Pro
۱	۰	0.00E+00	0
۲	۰	0.00E+00	0
۳	۰	0.00E+00	0
۴	۰	0.00E+00	0
۵	۱	2.03E-06	1.52E-06
۶	۱	2.44E-06	1.83E-06
۷	۰	0.00E+00	0
۸	۳	3.48E-06	2.61E-06
۹	۰	0.00E+00	0
۱۰	۱	2.78E-06	1.85E-06
۱۱	۴	2.88E-06	2.06E-06
۱۲	۲	6.75E-06	5.06E-06
۱۳	۴	3.39E-06	2.40E-06
۱۴	۴	4.29E-06	3.10E-06
۱۵	۵	3.28E-06	2.46E-06
۱۶	۱۴	3.87E-06	2.55E-06
۱۷	۱۲	3.74E-06	2.57E-06
۱۸	۱۵	3.61E-06	2.51E-06
۱۹	۱۰	4.23E-06	2.88E-06
۲۰	۱۷	4.92E-06	3.50E-06
۲۱	۱۸	7.42E-06	5.27E-06
۲۲	۷	7.07E-06	5.23E-06
X	۳۳	4.91E-06	3.68E-06
Y	۰	0.00E+00	0.00E+00
M	۰	0.00E+00	0.00E+00

همان‌طور که مشاهده می‌کنید بیشترین تعداد برهم‌کنش‌های کشف شده مربوط به کرموزوم X، بعد از آن مربوط به کروموزوم‌های ۱۶،۱۸،۲۰،۲۱ و کمترین آن مربوط به کروموزوم‌های ۱ تا ۹ در مجموعه داده مورد مطالعه می‌باشد. شکل ۳، میانگین مقدار تابع تناسب روش پیشنهادی در کشف برهم‌کنش‌های نواحی‌های پروموتر/انهنسر بالقوه به تفکیک هرکروموزوم در مقایسه با روش HiC-Pro در همان نواحی را نشان می‌دهد.

مطابق شکل ۳، همان‌طور که مشاهده می‌شود، مقادیر تابع تناسب برهم‌کنش‌های استخراج شده توسط روش پیشنهادی در اکثر کروموزوم‌ها بهتر از روش HiC-Pro می‌باشد. این بدین معنی است که، برهم‌کنش‌های استخراج شده توسط الگوریتم پیشنهادی در این نواحی از هر کروموزوم که بیان کننده پروموتر/انهنسر بالقوه می‌باشد، در مقایسه با روش HiC-Pro بهینه خواهد بود.



شکل ۳: میانگین مقدار تابع تناسب روش پیشنهادی و روش HiC-Pro به تفکیک هر کروموزوم

بحث و نتیجه‌گیری

امروزه، علم ژنتیک ثابت کرده است که، نواحی پروموتر/انهنسر در کروموزم‌های انسانی که با یکدیگر برهم‌کنش دارند، دارای طول ثابت (Fix-bin size) نمی‌باشند، بلکه طول آن‌ها می‌تواند متغیر باشد [۱۴]. همچنین تا به امروز، تحقیقات انجام شده [۵،۶،۱۵] قادر به کشف برهم‌کنش‌های بین نواحی ثابت ژنوم می‌باشند، که ممکن است ناحیه‌های کشف شده (پروموتر/انهسرهای بالقوه) دقیق نبوده و برهم‌کنش‌های شمارش شده بین آن‌ها صحیح نباشد.

در این پژوهش، یک الگوریتم فراابتکاری چند هدفه مبتنی بر شبیه‌سازی تبرید جهت حل مشکل فوق ارائه شد. الگوریتم پیشنهادی قادر به کشف پروموتر/انهنسرهای بالقوه با طول متغیر و شمارش برهم‌کنش‌های بین آن‌ها می‌باشد. نتایج و مقایسه‌های انجام گرفته نشان از کارایی بالا و بهتر بودن روش پیشنهادی در کشف پروموتر/انهنسر با طول متغیر نسبت به روش HiC-Pro است. در واقع روش HiC-Pro یکی از ابزارهای قدرتمند آنالیز داده‌های HiC می‌باشد، که قادر به کشف پروموتر/انهسر با طول متغیر نمی‌باشد.

روش پیشنهادی، قادر به کشف تعداد ۱۵۱ برهم‌کنش در تمامی کروموزوم‌ها می‌باشد، که روش HiC-Pro نمی‌تواند آن را کشف کند. برهم‌کنش‌های کشف شده حاصل از وجود برهم‌کنش‌های بین نواحی پروموتر/انهنسر بالقوه با طول متغیر می‌باشد که روش HiC-Pro قادر به کشف آن‌ها نیست. همچنین میانگین مقدار تابع تناسب روش پیشنهادی در کشف برهم‌کنش‌های نواحی‌های پروموتر/انهنسر بالقوه به تفکیک هر کروموزوم در مقایسه با روش HiC-Pro در همان نواحی را نشان می‌دهد. همان‌طور که مشاهده شد، مقادیر تابع تناسب برهم‌کنش‌های استخراج شده توسط روش پیشنهادی در اکثر کروموزوم‌ها بهتر از روش HiC-Pro می‌باشد. این بدین معنی است که، برهم‌کنش‌های استخراج شده توسط الگوریتم پیشنهادی در این نواحی از هر کروموزوم که بیان کننده پروموتر/انهنسر بالقوه می‌باشد در مقایسه با روش HiC-Pro بهینه خواهد بود.

نتایج به‌دست آمده نشان می‌دهند که، روش پیشنهادی نسبت به روش HiC-Pro موفق بوده است. روش HiC-Pro ، یکی از ابزارهای قدرتمند در زمینه آنالیز داده‌های Hi-C می‌باشد [۵]. کارایی این ابزار در آنالیز مجموعه داده‌های Hi-C از دیگر ابزارهای معرفی شده [۴-۷،۹-۱۲] که در بخش مقدمه ذکر گردید، بالاتر می‌باشد. همچنین روش پیشنهادی قادر به کشف و استخراج نواحی با طول متغیر از ژنوم است که این نواحی به صورت بالقوه می‌توانند پروموتر/انهنسر باشند. علاوه‌براین، روش پیشنهادی توانایی شمارش تمامی برهم‌کنش‌های بین این نواحی را دارد. تشخیص دادن ناحیه‌های پروموتر/انهنسر که با یکدیگر برهم‌کنش داشته باشند، مسئله مهمی است که به علم پزشکی در تشخیص زودهنگام سرطان کمک کرده و همچنین در درمان این بیماری می‌تواند مؤثر باشد؛ بنابراین پیشنهاد می‌شود، روش ارائه شده که قادر به استخراج صحیح این نواحی می‌باشد، بتواند جهت بهبود تشخیص بیماری سرطان به‌کار گرفته شود.

تشکر و قدردانی

مقاله مورد نظر مستخرج از پایان‌نامه دکترای تحت عنوان «کشف و استخراج برهم‌کنش‌های پروموتر/انهنسر و اجماعی از آن‌ها در ژنوم افراد سالم و ژنوم افراد سرطانی با استفاده از الگوریتم‌های فراابتکاری جدید» می‌باشد و از تمامی زحمات دکتر حمید علی‌نژاد رکنی عضو هیأت علمی دانشکده پزشکی دانشگاه ولز جنوبی سیدنی استرالیا به‌خاطر در اختیار گذاشتن

مجموعه داده Hi-C مربوط به سلول GM12878 نیز تشکر و قدردانی می‌گردد.

تضاد منافع

بدین‌وسیله نویسندگان تصریح می¬نمایند که هیچ‌گونه تضاد منافعی در خصوص پژوهش حاضر وجود ندارد.

References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin 2011;61(2):69-90.

2. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005;34(2):113-27.

3. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature methods. Nat Methods 2012;9(10):999-1003.

4. Hwang YC, Zheng Q, Gregory BD, Wang LS. High-throughput identification of long-range regulatory elements and their target promoters in the human genome. Nucleic Acids Res 2013;41(9):4835-46.

5. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol 2015;16:259.

6. Hwang YC, Lin CF, Valladares O, Malamon J, Kuksa PP, Zheng Q, et al. HIPPIE: a high-throughput identification pipeline for promoter interacting enhancer elements. Bioinformatics 2015;31(8):1290-2.

7. Wingett S, Ewels P, Furlan-Magaril M, Nagano T, Schoenfelder S, Fraser P, et al. HiCUP: pipeline for mapping and processing Hi-C data. F1000Res 2015;4:1310.

.۸ Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 2013;503(7475):290-4.

9. Cabreros I, Abbe E, Tsirig A. Detecting community structures in Hi-C genomic data. Annual Conference on Information Science and Systems (CISS); 2016 Mar 16-18; Princeton, NJ, USA: IEEE; 2016. p. 1-17.

10. Mifsud B, Tavares-Cadete F, Young AN, Sugar R. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet 2015;47(6):598-606.

11. Fotuhi Siahpirani A, Ay F, Roy S. A multi-task graph-clustering approach for chromosome conformation capture data sets identifies conserved modules of chromosomal interactions. Genome Biology 2016; 17:114.

12. Lazaris C, Kelly S, Ntziachristos P, Aifantis I, Tsirigos A. HiC-bench: comprehensive and reproducible Hi-C data analysis designed for parameter exploration and benchmarking. BMC Genomics 2017;18(1):22.

13. Arvey A, Agius P, Noble WS, Leslie CS. Sequence and chromatin determinants of cell-type–specific transcription factor binding. Genome Res 2012; 22(9):1723-34.

14. Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat Genet 2016; 48(5): 488–96.

15. Schmid MW, Grob S, Grossniklaus U. HiCdat: a fast and easy-to-use Hi-C data analysis tool. BMC Bioinformatics 2015;16:277.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2018; 5(2): 304-313

Detection and Extraction of Potential Promoter/Enhancer Interactions in Genome of Cancer Patients using an Evolutionary Multi-Objective Algorithm

Hosseinpoor Mohammadjavad[1], Parvin Hamid[2,6], Nejatian Samad[3,5]*, Rezaei Vahideh[4,5]

• **Received:** 22 Oct, 2017 • **Accepted:** 12 Jun, 2018

Introduction: Cancer, as one of the most common diseases, has influenced the health of many people. The main aim of this study was to present a multi-objective evolutionary algorithm. The algorithm is capable of detecting and extracting potentially promoter/enhancer areas in the chromosomes of the affected people using the information concerning inter-genomic interactions. The correct extraction of these areas can help early diagnosis of cancer.

Methods: In this applied and descriptive research, Hi-C data set including information on inter-genomic interactions in the GM12878 cell was used. Multi-objective evolutionary algorithm was used in order to discover and extract potential promoter /enhancer interactions. The mentioned algorithm was implemented using MATLAB software. Furthermore, the efficiency of this algorithm was evaluated using two criteria. The first criterion is a proportional function that calculates the magnitude of inter-genomic interactions relative to the length of the genome regions; and the second criterion is the number of discovered potential promoters/enhancers.

Results: The results and comparisons showed higher efficiency and optimality of the suggested method in discovering promoter/Enhancer interactions with variable length in comparison to HiC-Pro method. Therefore, the suggested method is able to discover the potential promoter/ enhancer interactions that cannot be discovered by HiC-Pro method.

Conclusion: The suggested algorithm is able to optimally discover and extract potential promoter/ enhancer with variable length. This is a great help in medical science for early diagnosis of cancer.

Keywords: Promoter, Enhancer, Hi-C Dataset, Multi-Objective Simulation Annealing Algorithm (MOSA), HiC- Pro Method

• **Citation:** Hosseinpoor MJ, Parvin H, Nejatian S, Rezaie V. Detection and Extraction of Potential Promoter/Enhancer Interactions in Genome of Cancer Patients using an Evolutionary Multi-Objective Algorithm. Journal of Health and Biomedical Informatics 2018; 5(2): 304-313.

1. Ph.D. student in computer engineering, Computer Engineering Dept., Yasooj Branch, Islamic Azad University, Yasooj, Iran
2. Ph.D. in Computer Engineering - Artificial Intelligence, Assistant Professor, Computer Engineering Dept., Nourabad Mamasani Branch, Islamic Azad University, Nourabad Mamasani, Iran
3. Ph.D. in Electrical Engineering, Assistant Professor, Electrical Engineering Dept., Yasooj Branch, Islamic Azad University, Yasooj, Iran
4. Ph.D. in Mathematics, Assistant Professor, Mathematics Dept., Yasooj Branch, Islamic Azad University, Yasooj, Iran
5. Member of Young Researchers and Elite Club, Yasooj Branch, Islamic Azad University, Yasooj, Iran.
6. Member of Young Researchers and Elite Club, Nourabad Mamasani Branch, Islamic Azad University, Nourabad Mamasani, Iran.

***Correspondence**: Electrical Engineering Dept., Yasooj Branch, Islamic Azad University, Yasooj, Iran.
• **Tel:** 09173414008 • **Email:** nejatian@iauyasooj.ac.ir