خوشه‌بندی مبتنی بر گراف با استفاده از آزمون ویلکاکسون جهت استخراج ارتباطات بیولوژیکی سلول‌ها و بافت‌ها

موسی مجرد [۱]، مربی؛ حمید پروین [۲]، استادیار؛ صمد نجاتیان [۳]، استادیار؛ وحیده رضایی [۴]، استادیار؛ کرم الله باقری فرد[۵]، استادیار

۱- دانشکده مهندسی برق و کامپیوتر – واحد یاسوج – دانشگاه آزاد اسلامی – یاسوج - ایران – m.mojarad@iauf.ac.ir

۲- دانشکده مهندسی کامپیوتر – واحد نورآباد ممسنی – دانشگاه آزاد اسلامی – نورآباد ممسنی – ایران-parvinhamid@gmail.com

۳- دانشکده مهندسی برق و کامپیوتر – واحد یاسوج – دانشگاه آزاد اسلامی – یاسوج – ایران-nejatian@iauyasooj.ac.ir

۴- دانشکده ریاضی – واحد یاسوج – دانشگاه آزاد اسلامی – یاسوج – ایران-v.rezaie@iauyasooj.ac.ir

۵- دانشکده مهندسی برق و کامپیوتر – واحد یاسوج – دانشگاه آزاد اسلامی – یاسوج – ایران-k.bagheri@iauyasooj.ac.ir

چکیده: شناسایی خوشه‌بندی مبتنی بر گراف یک روش کاربردی برای تشخیص ارتباط بین گره‌ها در شبکه‌های پیچیده بوده که نظرات قابل توجهی را به خود جلب کرده است. ازآنجایی‌که تشخیص جوامع مختلف در داده‌هایی با مقیاس بزرگ یک کار چالش‌برانگیز است، با درک ارتباط رفتار عناصر در جامعه(خوشه)، می‌توان ویژگی کلی خوشه‌ها را پیش‌بینی کرد. روش‌های خوشه‌بندی مبتنی بر گراف به دلیل توانایی آن‌ها برای نشان دادن ارتباط بین داده‌ها، نقش مهمی را در خوشه‌بندی داده‌های بیان ژن ایفا کرده‌اند. برای این‌که بتوان ژن‌های مؤثر در بروز بیماری‌ها را تشخیص داد، باید ارتباط بین سلول‌ها و یا بافت‌ها را به دست آورد. تعامل بین سلول‌ها و یا بافت‌های مختلف را می‌تواند با بیان ژن‌های مختلف بین آن‌ها نشان داد. در این پژوهش مسئله ارتباطات سلول به سلول و بافت به بافت به‌صورت یک گراف بیان‌شده و با تشخیص اجتماعات روابط استخراج می شوند. برای شبیه سازی و محا سبه میزان شباهت بین سلول‌ها و بافت‌ها از پایگاه داده فانتوم ۵ ا ستفاده می شود. پس از پیش‌پردازش و نرمال سازی داده‌ها، برای تبدیل این داده‌ها به گراف، میزان بیان ژن در سلول‌ها و بافت‌های مختلف بررسی‌شده و با در نظر گرفتن یک حد آستانه و آزمون ویلکاکسون، با استفاده از خوشه‌بندی ارتباطات بین آن‌ها شناسایی شدند.

واژه‌های کلیدی: خوشه‌بندی مبتنی بر گراف، بیان ژن، نرمال‌سازی، ویلکاکسون، ارتباطات سلول - سلول، ارتباطات بافت - بافت.

Graph-based Clustering using the Wilcoxon Test to Extract the Biological Communication of Cells and Tissues

Musa Mojarad[1], MSc.; Hamid Parvin[2], Assistant professor ; Samad Nejatian[3], Assistant professor; Vahideh Rezaie[4], Assistant professor; karamollah bagheifard[5], Assistant professor

1-Department of Computer Engineering, Yasooj Branch, Islamic Azad University, Yasooj, Iran, m.mojarad@iauf.ac.ir
2-Department of Computer Engineering, Nourabad Mamasani Branch, Islamic Azad University, Nourabad Mamasani, Iran, parvinhamid@gmail.com
3-Department of Electrical Engineering, Yasooj Branch, Islamic Azad University, Yasooj, Iran, nejatian@iauyasooj.ac.ir
4-Department of Mathematics, Yasooj Branch, Islamic Azad University, Yasooj, Iran, v.rezaie@iauyasooj.ac.ir
5-Department of Computer Engineering, Yasooj Branch, Islamic Azad University, Yasooj, Iran, m.mojarad@iauf.ac.ir

Abstract: Finding a graph-based clustering is an applied method for detecting the relationship between nodes in complex networks, which has attracted considerable attention. Since recognizing different communities in large-scale data is a challenging task, by understanding the relationship between the behavior of elements in a society (cluster), we can predict the general characteristics of the clusters. Graph-based clustering techniques have played an important role in the clustering of gene expression data due to their ability to show the relationship between data. In order to detect effective genes in the development of diseases, it is necessary to achieve the relationship between cells or tissues. The interaction between cells or different tissues can be demonstrated by expressing different genes between them. In this research, the problem of cell-to-cell and tissue-to-cell communication is expressed as a graph and is extracted by the recognition of relationships. The Phantom ۵ database is used to simulate and calculate the similarity between cells and tissues. After preprocessing and normalizing the data, for the conversion of these data to the graph, the expression of the gene in different cells and tissues has been examined and considering the threshold and the Wilcoxon test, using clustering of communications They were identified.

Keywords: Graph-based Clustering, Gene Expression, Normalization, Wilcoxon, Cell-Cell Communication, Tissue-Tissue Communication.

تاریخ ارسال مقاله: ۱۳۹۷/۰۵/۰۲
تاریخ اصلاح مقاله: ۱۳۹۷/۰۹/۱۹
تاریخ پذیرش مقاله: ۱۳۹۷/۱۱/۲۰
نام نویسنده مسئول: حمید پروین
نشانی نویسنده مسئول: دانشکده مهندسی برق و کامپیوتر – واحد نورآباد ممسنی– دانشگاه آزاد اسلامی – نورآباد – ایران.

۱- مقدمه

در دنیای واقعی، شبکه‌ها برای نشان دادن ارتباط بین انواع مختلفی از سیستم‌های پیچیده همانند شبکه‌های اجتماعی، شبکه‌های بیولوژیکی، شبکه‌های اینترنتی و غیره کاربرد دارند[۱]. یکی از ویژگی‌های برجسته شبکه‌ها که تبدیل به یک موضوع داغ تحقیقاتی شده، ساختار جامعه است[۲]. یک جامعه زیرمجموعه‌ای از گره‌ها است که وجه تشابه زیادی با یکدیگر دارند. ساختار جامعه نظیر ارتباطات بیان ژن‌ها، پروتئین‌ها و پروموترها، برای تشخیص انواع بیماری، تحلیل شبکه‌های اجتماعی و غیره کاربرد دارد. الگوریتم‌های مختلفی برای تشخیص جوامع طراحی‌شده‌اند[۴-۳]. این الگوریتم‌ها را می‌توان به‌طورکلی به‌صورت روش‌های راه رفتن تصادفی[۵]، خوشه‌بندی طیفی و پارتیشن‌بندی[۶]، مدولاسیون[۷]، ماتریس طیفی و فاکتورسازی ماتریس[۸]، تقسیم‌بندی دسته‌بندی کرد.

از جمله مواردی که می‌توان از شناسایی جامعه در علم پزشکی و بیوانفورماتیک استفاده کرد، تجزیه‌وتحلیل داده‌های بیان ژن می‌باشد[۹]. داده‌های بیان ژن بر اساس شرایط مختلف، برای تشخیص بیماری‌ها استفاده می‌شوند. با گسترش روزافزون فنّاوری، داده‌های زیادی از بیان ژن به‌دست‌آمده است، اما اطلاعات مفیدی از آن‌ها استخراج نشده است. به منظور استخراج اطلاعات از داده‌های بیان ژن می‌توان از شناسایی خوشه‌ها و اجتماعات بر اساس خوشه‌بندی مبتنی بر گراف استفاده کرد، به‌طوریکه هر اجتماع (خوشه) دارای بیشترین شباهت درونی است و با سایر اجتماعات بیشترین تفاوت را دارد[۱۰].

ادامه این تحقیق به شرح زیر است. مفهوم خوشه‌بندی بیان ژن و تحقیقات پیشین در بخش ۲ ارائه‌شده است. بخش ۳ در ارتباط با اصطلاحات فنی تحقیق (گراف‌کاوی، نرمال‌سازی، آزمون ویلکاکسون و شاخص سیلوئت) است. استخراج ارتباطات سلول به سلول و بافت به بافت مبتنی بر خوشه‌بندی بیان ژن در بخش ۴ مطرح می‌شود. بخش ۵ نتایج و آزمایش‌های مرتبط با الگوریتم خوشه‌بندی پیشنهادی گزارش شده است. در این بخش همچنین نتایج خوشه‌بندی و ارتباطات بین سلولی و بین بافتی روی مجموعه داده فانتوم۵ ارائه‌شده است. درنهایت نتیجه‌گیری و بررسی‌های بیشتر در بخش ۵ بحث می‌شود.

۲- پیشینه تحقیق

تحقیقات محاسباتی زیادی در زمینه ریاضی و گراف برای تحلیل رفتار سلول و بافت صورت گرفته که سهم گسترده‌ای به دانش بیولوژیکی[۱۱]، شناسایی اجتماعات[۱۲] و همچنین خوشه‌بندی[۱۳][۱۴] ارائه شده‌اند. از شناسایی اجتماعات به منظور بررسی ویژگی ساختاری گراف‌های پیچیده استفاده می‌شود[۱۵]. برای مثال در تحقیقات مختلف نشان داده شده است که کشف اجتماعات می‌تواند به شناسایی نویسندگان مقالات با موضوعات مشابه در گراف‌های اجتماعی کمک کند[۱۶].

در سال‌های اخیر برای تشخیص اجتماعات روش‌های متعددی پیشنهادشده است. یکی از الگوریتم‌های اولیه برای این منظور، الگوریتم برش کمینه می‌باشد[۱۷]. در الگوریتم برش کمینه، گراف به تعداد بخش‌های از قبل تعیین‌شده تقسیم‌شده به‌طوری‌که تعداد یال‌هایی که بین اجتماعات قرار دارند مینیمم می‌باشد. الگوریتم‌های افراز گراف نظیر الگوریتم‌های خوشه‌بندی طیفی[۱۸] دسته دیگری از الگوریتم‌های شناسایی اجتماعات می‌باشند. همچنین الگوریتم‌های مبتنی بر بهینه‌سازی، شامل بهینه‌سازی ماژولاریتی، بهینه‌سازی خارجی، بهینه‌سازی طیفی برای حل مسئله کشف اجتماعات پیشنهادشده است[۱۹،۱۶].

از دیگر الگوریتم‌ها که به‌صورت گسترده برای شناسایی اجتماعات استفاده می شوند، الگوریتم‌های خوشه‌بندی سلسله مراتبی [۲۰] هستند که از معیار شباهت بین جفت گره‌ها برای خوشه‌بندی استفاده می‌کنند. در این الگوریتم‌ها گره‌های با بیشترین معیار تشابه در یک اجتماع قرار می‌گیرند. نیومن، الگوریتمی تقسیمی مبتنی بر گراف دوبخشی سلسله‌مراتبی را برای شناسایی اجتماعات پیشنهاد داده است[۲۱]. در این الگوریتم یالی که دارای مقدار دوبخشی بالایی می‌باشد حذف می‌گردد. نیومن مسئله کشف اجتماعات را به‌عنوان یک مسئله بهینه‌سازی با در نظر گرفتن معیاری بنام ماژولاریتی Q بیان نموده است.

در این تحقیق از یک روش خوشه‌بندی مبتنی بر گراف براساس پیمانگی برای خوشه‌بندی داده‌های بیان ژن استفاده می‌کنیم. به منظور کاهش پراکندگی داده‌های بیان ژن و سازگاری آنها با الگوریتم‌های خوشه‌بندی پایه نیاز به نرمال‌سازی و محاسبه ارتباطات معنایی بین نمونه‌ها می‌باشد. هدف از خوشه‌بندی داده‌های بیان ژن، استخراج ارتباطات بین سلول‌ها و بین بافت‌های مختلف در این داده‌ها می‌باشد. با توجه به خصوصیات منحصر به فرد داده‌های بیان ژن و تفاوت آن با مجموعه داده‌های عمومی، روش پیشنهادی با نگاشت این داده‌ها به گراف علاوه بر کاهش حجم محاسباتی، امکان اعمال الگوریتم‌های خوشه‌بندی پایه مبتنی بر گراف را بر روی این داده‌ها فراهم می‌کند. بنابراین این مورد را می‌توان به‌عنوان یکی از موارد اصلی تفاوت روش پیشنهادی با سایر روش‌های مشابه بیان نمود.

۳- اصطلاحات فنی تحقیق

در این بخش اصطلاحات فنی تحقیق شامل گراف‌کاوی، نرمال سازی و آزمون ویلکاکسون بررسی می‌شوند. گراف‌کاوی یکی از روش‌های جدید برای استخراج داده‌هایی است که توسط گراف نشان داده می شوند. نرمال‌سازی یکی از گام‌های ضروری برای تحلیل داده‌های بیان ژن است و آزمون ویلکاکسون از دسته آزمون‌های آماری جهت ارزیابی ارتباطات معنایی دو نمونه وابسته با مقیاس رتبه‌ای می‌باشد.

۳-۱- گراف‌کاوی

گراف G یک مجموعه دوتایی به‌صورت G = (V, E) است که V مجموعه‌ای از گره‌ها و E شامل یال‌های گراف می‌باشد. $n = |V|$ تعداد گره‌ها (مرتبه گراف) و $m = |E|$ تعداد یال‌ها (اندازه گراف) را نشان می‌دهد. در یک گراف وزن‌دار، تابع وزن W به صورت $W \rightarrow R$ تعریف‌شده که وزنی را به هر یال از گراف اختصاص می‌دهد. به‌طورکلی چگالی یک گراف نسبت تعداد یال‌های موجود در گراف به حداکثر یال‌های ممکن می‌باشد. رابطه (۱) چگالی یک گراف را تعریف می کند[۲۲].

$$\partial(G) = \frac{m}{n} for\ n \in \{۰,۱\}, we\ set\ \partial(G) = ۰ \qquad (۱)$$

معیارهای مختلفی برای تحلیل گراف باتوجه‌به اندازه آن وجود دارد که می‌توان به مرکزیت نزدیکی، شاخص پیوستگی، قدرت، مرکزیت بردار مشخص، دسترسی، پیوستگی و درجه گره اشاره کرد[۲۳].

۳-۲- نرمال‌سازی داده‌های بیان ژن

نرمال‌سازی یک گام ضروری با تأثیر قابل‌توجه در تحلیل داده‌های بین ژن می‌باشد. در زمینه بیان ژن، RNA-seq یک نوع فنّاوری بوده که با بهره‌گیری از تکنیک «توالی‌یابی نسل بعدی» برای به دست آوردن تصویری کلی از مقدار ژنوم RNA در یک بازه زمانی خاص استفاده می‌شود[۲۴]. روش‌های نرمال‌سازی RNA-seq مقایسه تفاوت بیان ژن بین نمونه‌ها را فراهم می‌کنند. بنابراین برای تعیین مناسب‌ترین روش‌های نرمال‌سازی داده‌های بیان ژن، در ادامه مقایسه‌ای بین برخی از روش‌های معمول در این زمینه انجام می‌شود.

- روش‌های نرمال‌سازی درون نمونه و بین نمونه: این نوع روش‌ها اصلاح سطح بیان در هر ژن مرتبط با ژن‌های دیگر در همان نمونه را نشان می‌دهد. از رایج‌ترین روش‌ها در این زمینه RPKM[۲۵] و FPKM[۲۶] می‌باشند.
- روش‌های نرمال‌سازی درون نمونه: تغییرات موجود در شمارش خواندن یک ژن بین نمونه‌ها به دلیل تفاوت در عمق توالی است، به همین دلیل نرمال‌سازی درون نمونه از خواندن‌های خام استفاده می‌کند. ساده‌ترین روش نرمال‌سازی در این زمینه، TC می‌باشد.
- روش‌های نرمال‌سازی سراسری: ازآنجاکه تنوع در میان ژن‌های یک نمونه و تغییرات هر ژن در سراسر نمونه‌ها باید اصلاح شود، دو روش Med-pgQ2 و UQ-pgQ2 معرفی شدند[۲۴]. در Med-pgQ2 شمارش نرمال شده جدید $Y_{gj}^{Med-pgQ2}$ برای هر ژن و هر ۱۰۰ خواندن به‌صورت رابطه (۲) تعریف می‌شود.

$$Y_{gj}^{Med-pgQ2} = \frac{Y_{gj}^{Med}}{Q2_g^{Med}} \times 100 \qquad (۲)$$

در این رابطه Y_{gj}^{Med} مقادیر عبارت برای ژن g در نمونه j و همچنین $Q2_g^{Med}$، میانه ژن g پس از نرمال‌سازی هر نمونه می‌باشد.

در UQ-pgQ2 فرض می‌شود Y_{gj}^{UQ} مقادیر عبارت برای ژن g در نمونه j و نرمال شده توسط UQ (۷۵ درصد) است؛ همچنین فرض می‌شود $Q2_g^{UQ}$، میانه ژن g در طول نمونه‌ها بعد از UQ نرمال‌سازی شده است. بنابراین، شمارش نرمال شده جدید $Y_{gj}^{UQ-pgQ2}$ برای هر ژن و هر ۱۰۰ خواندن به‌صورت رابطه (۳) تعریف می‌شود.

$$Y_{gj}^{UQ-pgQ2} = \frac{Y_{gj}^{UQ}}{Q2_g^{UQ}} \times 100 \qquad (۳)$$

- روش‌های نرمال‌سازی مقیاس‌پذیر: این روش‌ها برای محاسبه مقیاس‌های پوششی و همچنین جهت نرمال‌سازی مقادیر بیان ژن سلول‌ها/بافت‌ها با ابعاد بسیار بالا بکار برده می‌شوند. از رایج‌ترین روش‌ها در این زمینه RLE و TMM می‌باشند[۲۷]. در RLE، ابتدا برای هر ژن i، میانگین برای هر نمونه محاسبه می‌شود، یعنی $med(y_{*j})$. به‌طوریکه y_{*j} مقدار j-مین ستون از ماتریس $[y_{ij}]$ است. سپس مقدار انحراف از میانگین $(y_{ij} - med(y_{*j}))$ محاسبه می‌شود، به‌طوریکه y_{ij} لگاریتم بیان‌شده هر ژن j از نمونه i-ام می‌باشد.

به منظور ارزیابی عملکرد روش‌های نرمال‌سازی RNA-seq معرفی شده، در ادامه تحلیل و ارزیابی تفاوت ژن‌های بیان‌شده با استفاده از مقادیر AUC بررسی می‌شوند. جدول ۱ مقدار AUC را برای روش‌های مختلف با توجه به آزمون z-test دو نمونه‌ای یک‌طرفه نشان می‌دهد.

جدول ۱: مقایسه روش‌های نرمال‌سازی مختلف

Methods	z-statistics	p-value
RLE	۰/۷۲۱۵	۰/۲۳۲۵
UQ-pgQ2	۰/۷۵۵۴	۰/۲۲۵۰
FPKM	۲/۰۰۸۲	۰/۰۲۲۳
TMM	۲/۰۰۹۶	۰/۰۲۲۲۴
DESeq	۲/۵۸۲۶	۰/۰۰۴۹
FQ	۲/۶۸۶۱	۰/۰۰۳۶
TC	۲/۷۵۱۷	۰/۰۰۳۰

در این جدول، به ازای هر روش، فهرستی از نتایج مقادیر p بر روی مجموعه داده فانتوم ۵ ارائه شده است[۲۸]. نتایج به دست آمده نشان‌دهنده برتری مقدار AUC در روش RLE نسبت به سایر روش‌ها می‌باشد.

۳-۳- آزمون ویلکاکسون

آزمون ویلکاکسون[۲۹] از آزمون‌های آماری نا پارامتری است که برای ارزیابی همانندی دو نمونه وابسته با مقیاس رتبه‌ای به کار می‌رود. آزمون‌های نا پارامتری علامت، مک‌نمار و ویلکاکسون، برای مقایسه و کشف ارتباطات معنایی زوجی بین نمونه‌ها استفاده می‌شوند. آزمون علامت و ویلکاکسون، نیاز به متغیر با سطوح زیاد دارند. آزمون ویلکاکسون یک مزیت نسبت به آزمون علامت دارد و آن نمایش شدت تفاوت بین دو نمونه است. آزمون مک‌نمار برای متغیرهایی که دو سطحی کاربرد دارد و برای متغیرهای بیش از دو سطح، قابل انجام نمی‌باشد. بنابراین، ازآنجایی‌که هدف این تحقیق بررسی تفاوت معنایی چند متغیر

است، از آزمون ویلکاکسون استفاده می‌شود. ازجمله الزامات اجرای این آزمون، مرتبط بودن داده‌ها و همچنین دارای بودن مقیاس مرتبه‌ای برای مقادیر جفت‌ها است تا بتوان تفاوت درون جفت‌ها را محاسبه کرد.

برای محاسبه آزمون ویلکاکسون فرض می‌کنیم نمونه‌ای با ابعاد n در اختیار است که داده‌های آن به‌صورت جفتی هستند؛ بنابراین ۲n داده در دسترس می‌باشد. برای جفت‌های $X_{1,i}$ و $X_{2,i}$ به‌طوریکه $i = 1,2,\dots,n$ فرض‌های صفر و یک مطرح می‌شود. فرض صفر؛ تفاوت بین جفت‌ها دارای توزیع متقارن حول صفر و فرض یک؛ تفاوت بین جفت‌ها دارای توزیع متقارن حول صفر را بیان می‌کند. دراینجا به ازای همه اندازه‌های جفتی، مقدار $|X_{2,i} - X_{1,i}|$ محاسبه‌شده و علامت تفاضل نیز ثبت می‌شود. در مرحله بعد تمامی تفاضل‌های صفر را حذف کرده و ابعاد نمونه جدید را n_r می‌نامیم. سپس داده‌ها را در نمونه جدید از کوچک‌ترین تا بزرگ‌ترین قدر مطلق تفاضل مرتب کرده و به داده‌ها رتبه داده می‌شود (کوچک‌ترین رتبه یک). این رتبه با متغیر R_i نشان داده می‌شود. درنهایت آزمون ویلکاکسون مطابق رابطه (۴) تعریف می‌شود.

$$W = \sum_{i=1}^{n_r} \left[sign(X_{2,i} - X_{1,i}) \times R_i \right] \quad (۴)$$

فرض صفر را درصورتیکه $|W| > W_{critical,n_r}$ باشد، رد می‌شود.

۳-۴- شاخص سیلوئت

با توجه به نبود کلاس هدف در داده‌های بیان زن، نیاز به شاخص‌های اعتبارسنجی درونی برای سنجش میزان صحت نتایج خوشه‌بندی می‌باشد. در این تحقیق از شاخص درونی سیلوئت [۳۰] برای این منظور استفاده‌شده است. این معیار عمل ارزیابی خوشه‌ها را با استفاده از مقادیر درونی هر خوشه و نمای آن‌ها محاسبه می‌کند.

شاخص سیلوئت بر مبنای محاسبه اعتبار خوشه بر اساس تفاوت دوبه‌دوی فاصله‌ی بین و درون خوشه‌ای می‌باشد و ترکیبی از شباهت درون خوشه‌ای و بین خوشه‌ای ارائه می‌دهد. میانگین این شاخص می‌تواند مقداری در بازه [۱+،۱-] اختیار کند. اگر میانگین شاخص نزدیک به عدد ۱ باشد آنگاه مدل خوشه‌بندی رضایت‌بخش تلقی می‌شود. مقادیر منفی و نزدیک به صفر این شاخص حاکی از نامناسب بودن مدل و عملکرد ضعیف الگوریتم خوشه‌بندی در ایجاد خوشه‌ها است. محاسبه این شاخص برای یک نمونه داده مانند x_i در سه‌گام انجام می‌شود.

گام ۱: محاسبه متوسط فاصله داده x_i از تمام داده‌های دیگر در خوشه خودش (a_i).

گام ۲: محاسبه متوسط فاصله داده x_i از تمام داده‌های دیگر در خوشه دیگر. کمترین مقدار به‌دست‌آمده از بین K-۱ خوشه، متوسط فاصله محاسبه‌شده را برای انتخاب مشخص می‌کند (b_i).

گام ۳: ضریب سیلوئت با رابطه (۵) محاسبه شود.

$$S_i = \frac{b_i - a_i}{\max(b_i, a_i)} \quad (۵)$$

در این رابطه a_i و b_i به ترتیب بیانگر میانگین فاصله بین مشاهده i با سایر مشاهدات در یک خوشه مشابه و میانگین فاصله مشاهده i به تمام مشاهدات در خوشه‌های دیگر می‌باشد. به‌منظور بررسی مناسب بودن یک روش خوشه‌بندی، متوسط S_i برای تمام داده‌ها محاسبه می‌شود.

۴- استخراج ارتباطات بین سلولی و بین بافتی

روش‌های خوشه‌بندی بیان ژن اجازه می‌دهد تا هزاران ژن و یا حتی بیشتر در دسته‌های کوچک‌تر قرار گیرند. یکی از ویژگی خوشه‌بندی بیان ژن، تعریف اندازه‌گیری شباهت (برای مثال فاصله) بین مشخصات بیان ژن است[۳۱]. در این پژوهش برای به دست آوردن ارتباط بین سلولی/بافتی از روش ویلکاکسون استفاده می‌شود که اینکار براساس داده‌های بیان ژن انجام می‌شود.

یک روش خوشه‌بندی مناسب نقش کلیدی دریافتن ارتباطات بین سلول‌ها/بافت‌ها به عهده دارد. در کار قبلی تشخیص ارتباط‌های بین سلول‌ها/بافت‌ها در بیماری‌های مختلف با توجه به مشخصات ساختار توپولوژیکی گراف و یک روش خوشه‌بندی تجمعی بهبودیافته انجام شد. روش پیشین دارای دو مرحله بود؛ در مرحله اول چندین مدل خوشه‌بندی به‌منظور تشخیص ارتباط‌های اولیه بین سلول‌ها/بافت‌ها در جهت تولید نتایج بهتر نسبت به الگوریتم‌های انفرادی، ترکیب می‌شدند و در مرحله دوم تشابه بین سلول‌ها/بافت‌ها در هر خوشه با استفاده از یک معیار شباهت مبتنی بر ساختار توپولوژیکی گراف محاسبه می‌شد. در این تحقیق از کارایی یک الگوریتم مبتنی بر گراف برای تشخیص ارتباط‌های بین سلولی/بافتی استخراج شده توسط آزمون ویلکاکسون استفاده می‌شود.

با توجه به پراکندگی مقادیر پایگاه داده فانتوم ۵ (اطلاعات مربوط به این دیتاست در آدرس http://fantom.gsc.riken.jp/5 موجود می باشد)، ابتدا بایستی داده‌های موردنظر را نرمال‌سازی کنیم. هدف از نرمال‌سازی جلوگیری از پراکندگی داده‌ها با قرار دادن مقادیر داده‌ها در یک بازه مشخص است. روش‌ها و متدهای مختلفی جهت نرمال سازی وجود دارد، در [۳۲] از روش RLE جهت نرمال سازی داده‌هایی مشابه مجموعه داده فانتوم ۵ استفاده و نتایج خوبی گزارش‌شده است. در این تحقیق نیز از این روش جهت نرمال‌سازی بهره گرفته می‌شود.

در مجموعه داده فانتوم ۵ سطرها بیانگر شماره بیان ژن‌ها و ستون ها معرف نمونه‌های مختلف سلول/ بافت می‌باشند که از نمونه‌های انسانی متعددی استخراج شده است، به‌طوری‌که از یک سلول/ بافت ممکن است چند نمونه وجود داشته باشد. جدول ۲ شمایی از مجموعه داده فانتوم ۵ را با ۱۰۸ نمونه و ۸۶۴۲۸ ژن برای سلول‌ها نشان می‌دهد.

بعد از نرمال‌سازی داده‌ها باید ماتریس همبستگی بین ستون‌ها جهت تشخیص ارتباطات بین سلولی/ بافتی محاسبه شود. برای به دست آوردن ماتریس همبستگی از هر سلول/ بافت حداقل ۲ نمونه مورد نیاز است، لذا نمونه‌هایی با یک‌بار مشاهده در مجموعه داده، در نظر گرفته نمی شوند. برای به دست آوردن ارتباط بین سلولی/بافتی از روش ویلکاکسون و با مقدار p-value بزرگ‌تر از ۰/۰۵ استفاده‌شده

است. خروجی ماتریس همبستگی تعداد بیان ژن‌های مشترک بین هر دو نمونه سلول/بافت را نشان می‌دهد. به طور مشابه جدول ۳ شنایی از مجموعه داده فانتوم ۵ را با ۴۰ نمونه و ۸۶۴۲۸ ژن برای بافت‌ها نشان می‌دهد.

جدول ۲: شمایی از مجموعه داده فانتوم ۵ برای سلول‌ها.

	سلول ۱۰۸			...	سلول ۱			
	نمونه شماره N	نمونه شماره ۲	نمونه شماره ۱	⋮	نمونه شماره N	نمونه شماره ۲	نمونه شماره ۱	
مقدار بیان ژن	...	...	...	...	...	...	مقدار بیان ژن	ژن شماره ۱
	...	...	...	...	...	...	...	...
مقدار بیان ژن	...	...	...	...	...	...	مقدار بیان ژن	ژن شماره ۸۶۴۲۸

جدول ۳: شمایی از مجموعه داده فانتوم ۵ برای بافت‌ها.

	بافت ۴۰			...	بافت ۱			
	نمونه شماره N	نمونه شماره ۲	نمونه شماره ۱	⋮	نمونه شماره N	نمونه شماره ۲	نمونه شماره ۱	
مقدار بیان ژن	...	...	...	...	...	...	مقدار بیان ژن	ژن شماره ۱
	...	...	...	...	...	...	...	...
مقدار بیان ژن	...	...	...	...	...	...	مقدار بیان ژن	ژن شماره ۸۶۴۲۸

درواقع با روش ویلکاکسون و مقدار p-value ارتباط معنایی هر جفت سلول/بافت برای تمام ژن‌ها محاسبه می‌شود. جدول ۴ شمایی از خروجی ماتریس همبستگی را برای سلول‌ها و جدول ۵ این اطلاعات را برای بافت‌ها نشان می‌دهد.

برای مثال اگر سلول/بافت شماره ۱ با سلول/بافت شماره ۲ در ۵۰ بیان ژن، مشترک باشد، مقدار درایه ماتریس برای این دو سلول/ بافت (تعداد ژن‌های بیان شده مربوطه) برابر ۵۰ می‌باشد. به‌طورکلی جداول ۴ و ۵ به ترتیب گراف‌های ارتباطات سلول به سلول و بافت به بافت را تشکیل می‌دهند به‌طوری‌که گره‌ها بیانگر سلول‌ها/بافت‌ها و یال‌ها بیانگر وزن‌های بین آنها می‌باشند.

جدول ۴: خروجی ماتریس همبستگی برای سلول‌ها.

سلول ۱۰۸	...	سلول ۲	سلول ۱	
تعداد ژن بیان‌شده مربوطه		تعداد ژن بیان‌شده مربوطه	۰	سلول ۱
...		۰	...	سلول ۲
...	۰	...	...	...
۰	...	تعداد ژن بیان‌شده مربوطه	تعداد ژن بیان‌شده مربوطه	سلول ۱۰۸

جدول ۵: خروجی ماتریس همبستگی برای بافت‌ها.

بافت ۴۰	...	بافت ۲	بافت ۱	
تعداد ژن بیان‌شده مربوطه		تعداد ژن بیان‌شده مربوطه	۰	بافت ۱
...		۰	...	بافت ۲
...	۰	...	...	...
۰	...	تعداد ژن بیان‌شده مربوطه	تعداد ژن بیان‌شده مربوطه	بافت ۴۰

با توجه به اینکه گراف ایجادشده در بخش بازنمایی کامل است، اما واضح است که وزن همه یال‌ها تأثیر در خوشه‌بندی سلول‌ها/ بافت‌ها ندارند و وجود آن‌ها ممکن است باعث افت کارایی روش پیشنهادی شود. به این منظور، قبل از اجرای الگوریتم تشخیص جوامع، با اعمال یک آستانه بر روی یال‌های گراف، آن دسته از یال‌ها را که وزنشان از آستانه θ کمتر باشد را حذف می‌کنیم.

در مرحله بعد به‌منظور تشخیص ارتباطات بین سلولی و بین بافتی از یک روش خوشه‌بندی مبتنی بر گراف استفاده می‌کنیم. با اینکار سلول‌ها/بافت‌ها را در یک اجتماع (یا خوشه) قرار می‌دهیم و ارتباطات را استخراج می‌کنیم. به‌طوری‌که با استفاده از این ارتباطات می‌توان بررسی کرد که؛ اولاً در کدام ارتباط‌ها ژن‌ها به‌صورت یکسان بیان‌شده‌اند، ثانیاً چه بیان ژن‌هایی در چندین سلول/بافت به‌صورت یکسان می‌باشند و ثالثاً در یک سلول/بافت خاص چه بیان ژن‌هایی وجود دارد.

هدف خوشه‌بندی در اینجا قرار گرفتن سلول‌ها/بافت‌های مشابه در دسته‌های یکسانی است. اکثر روش‌های خوشه‌بندی گراف دارای مشکلات و معایبی هستند[۳۳] در اغلب روش‌ها پارامتر k (تعداد خوشه‌ها)، باید توسط کاربر و قبل از اجرای الگوریتم مشخص شود. از طرفی توزیع داده‌ها در هر خوشه، از معیارهای مهم در خوشه‌بندی است که در اکثر روش‌های گذشته در نظر گرفته نشده است. در نظر گرفتن میزان پراکندگی ویژگی های موجود در هر خوشه، عملکرد الگوریتم خوشه‌بندی را بالا می‌برد. در این تحقیق، برای حل این مسئله و برطرف کردن مشکل فوق، از الگوریتم خوشه‌بندی مبتنی بر گراف Louvain [۴] استفاده می‌شود. Louvain یک الگوریتم الگوریتم حریصانه است که سعی در بیشینه‌سازی معیار پیمانگی در گراف دارد. برخلاف تمامی روش‌های خوشه‌بندی دیگر، در این روش محدودیت اندازه گراف ورودی به محدودیت حافظه و همچنین زمان پردازش بستگی ندارد. به همین دلیل، این الگوریتم به‌راحتی بر گراف‌هایی با صدها میلیون گره و به‌صورت توزیع‌شده قابل‌اعمال است.

بهینه‌سازی تابع پیمانگی یک روش برای تشخیص جوامع است که به‌صورت گسترده استفاده می‌شود. پیمانگی در الگوریتم فوق ارزش بهینه سازی شده‌ای بین ۱- و ۱ تعریف می‌شود که چگالی لینک‌ها درون جوامع را نسبت به ارتباط بین جوامع اندازه‌گیری می‌کند. رابطه (۶) تابع پیمانگی را نشان می‌دهد.

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \qquad (۶)$$

در این رابطه A_{ij} وزن بین دو گره i و j، k_i و k_j مجموع وزن‌های یال‌های متصل به گره‌های i و j است. m مجموع تمام وزن یال‌ها در گراف است؛ c_i و c_j ارتباطات گره‌های i و j است. δ نیز یک تابع دلتا ساده مطابق رابطه (۷) است.

$$\delta_{ij} = \begin{cases} 0 & if\ i \neq j \\ 1 & if\ i = j \end{cases} \quad (۷)$$

بهینه‌سازی کامل تابع پیمانگی با الگوریتم خوشه‌بندی Louvain در دو مرحله انجام می‌شود:

۱- با بهینه‌سازی محلی، این الگوریتم به دنبال گروه‌های کوچک می‌گردد. برای گره i منفعت تخصیص به خوشه C با استفاده از رابطه (۸) محاسبه می‌شود.

$$AQ = \left[\frac{\sum_{in}+k_{i,in}}{2m} - \left(\frac{\sum_{tot}+k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}+k_i}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \quad (۸)$$

که در آن $\sum_{in}$ مجموع وزن‌ها در خوشه C است، $\sum_{tot}$ مجموع وزن یال‌هایی است که به گره‌های خوشه C وصل می‌شود، k_i مجموع یال‌های گره i و $k_{i,in}$ نشان‌دهنده مجموع وزن یال‌هایی است که یک سر آن گره i است و یکسر دیگر آن نیز خوشه C است. همچنین m برابر مجموع وزن تمام یال‌های گراف است.

۲- سپس با ادغام گروه‌های کوچک که توانایی ایجاد گروه‌های بزرگ‌تر را دارند خوشه‌بندی را ادامه می‌دهد. مراحل تا جایی تکرار می‌شوند تا تغییری در خوشه‌ها به دست نیاید و پیمانگی به حالت بیشینه برسد.

با توجه به ادغام خوشه‌های کوچک به شرط افزایش مقدار پیمانگی الگوریتم Louvain به صورت خودکار تعداد خوشه‌های نهایی را تشخیص می‌دهد. بنابراین تعداد خوشه‌های نهایی برابر تعداد تعداد خوشه‌ها با مقدار پیمانگی بیشینه می‌باشد. شکل ۱، شبه کد روش پیشنهادی برای استخراج ارتباطات بین سلولی/ بین بافتی را نشان می‌دهد.

Cell-to-Cell Communication Extraction Algorithm
Input: FANTOM5 dataset
Output: Cells/Tissues clustering

1: Apply filters to dataset (Remove cells with only one sample and Lines without GeneID)
2: Use Relative Log Expression (RLE) to normalize FANTOM5 dataset.
3: for i = 1 to nCells/nTissues do
4: for j = 1 to nCells/nTissues do
5: if $i \neq j$ then
6: GraphCommunication (i,j) = Use the Wilcoxon test to calculate Communication between i and j cells/ Tissues, with p-value > 0.05
7: else
8: GraphCommunication(i,j) = 0
9: end if
10: end for
11: end for
12: Apply LouvainAlgorithm(GraphCommunication) clustering method to detect Communication with the highest Gene Expression.

شکل ۱ : شبه کد الگوریتم پیشنهادی

در خط ۱، فرایند فیلترینگ مجموعه داده فانتوم ۵ انجام می‌شود و سلول/بافت‌هایی با تنها یک نمونه و همچنین بیان ژن‌هایی بدون GeneID حذف می‌گردند. در خط ۲، مجموعه داده فانتوم ۵ با استفاده از روش RLE نرمال می‌شود. در خط ۳، متغیر i به تعداد سلول‌ها یا بافت‌ها تکرار می‌شود. خط ۴، متغیر j نیز به صورت مشابه به تعداد سلول‌ها یا بافت‌ها تکرار می‌شود. این کار بدین منظور است که ارتباط هر جفت سلول را شناسایی کنیم. خط ۵، نشان می‌دهد که آیا دو سلول i و j شبیه هم هستند یا خیر. اگر یکسان نباشند در خط ۶، ارتباط معنایی بین آنها با استفاده از تکنیک آماری ویلکاکسون محاسبه می‌شود. در اینجا ارتباطاتی با مقدار p-value >0.05 در نظر گرفته می‌شوند. خط ۷و ۸ نشان می‌دهد که دو سلول/بافت شبیه هم دارای ارتباط ۰ می‌باشد. خطوط ۹، ۱۰ و ۱۱ پایان حلقه‌های تکرار می‌باشد. در خط ۱۲ الگوریتم خوشه‌بندی Louvain بر مبنای p-value >0.05 اعمال شده و سلول‌ها/ بافت‌ها را خوشه‌بندی می‌کند.

۵- نتایج و آزمایش‌ها

در این بخش عملکرد الگوریتم پیشنهادی به‌منظور کشف ارتباطات بین سلول‌ها/بافت‌ها روی مجموعه داده واقعی فانتوم ۵ مورد ارزیابی قرار می‌گیرد. در داده‌های فانتوم ۱۸۲۹ نمونه با ۲۰۱۸۰۲ پروموتر (بیان ژن) وجود دارد. ویژگی‌ها در مجموعه داده فانتوم پروموترها هستند که درواقع حاوی اطلاعاتی در مورد ژن‌ها می‌باشد که منجر به تولید آن شده است. هدف از استخراج این ارتباطات درواقع شناسایی سلول‌ها یا بافت‌هایی است که در یک یا چند بیماری بیان ژن مشابهی دارند.

در این مجموعه داده نمونه‌ها (ستون‌ها) شامل نام سلول‌ها و بافت هایی از بیماران مختلف می‌باشند. پروموتر ها (سطرها) بیانگر شماره بیان ژن‌ها ا ست که با ا ستفاده از EntreGene_ID م شخص می شود. بعضی از مقادیر EntreGene _ID در پایگاه داده اصلی بدون مقدار هستند و با NA مشخص شده‌اند، لذا این سطرها حذف می‌گردند. ازآنجایی که هدف ما به دست آوردن ارتباط بین سلولی/بافتی است، لذا فقط ستون‌های که مربوط به سلول یا بافت هستند، در نظر گرفته می‌شوند. در اینجا ۷۰۲ ستون مربوط به سلول‌ها و ۱۲۵ ستون مربوط به بافت‌ها مشخص شدند. به‌طورخاص از یک سلول/بافت ممکن ا ست چند نمونه گرفته شده با شد. جدول ۶ خلا صه ای از اطلاعات مجموعه داده فانتوم ۵ را نشان می‌دهد.

جدول ۶: خلاصه‌ای از اطلاعات مجموعه داده فانتوم ۵.

	تعداد نمونه‌ها	تعداد پروموترها	تعداد سلول‌ها	تعداد بافت‌ها
پایگاه داده واقعی	۱۸۲۹	۲۰۱۸۰۲	۴۹۸	۱۹۴
بعد از فیلتر شدن	سلول: ۱۲۵ بافت: ۷۰۲	۸۶۴۲۷	۱۰۸	۴۰

با توجه به نبود کلاس هدف، نیاز به رو شاخص‌های اعتبارسنجی درونی برای سنجش میزان صحت نتایج خوشه‌بندی داریم. هدف از اعتبارسنجی خوشه‌ها یافتن خوشه‌هایی است که بهترین تناسب را با داده‌های موردنظر داشته باشند. داده‌های متعلق به یک خوشه بایستی تا

حد ممکن به یکدیگر نزدیک باشند(معیار تراکم). معیار رایج برای تعیین میزان تراکم داده‌ها، واریانس داده‌ها است. همچنین خوشه‌ها خود بایستی به‌اندازه کافی از یکدیگر جدا باشند(معیار جدایی). در این تحقیق از شاخص سیلوئت برای این منظور استفاده‌شده است.

ارتباط دو سلول/بافت زمانی به وجود می‌آید که در تعدادی از سلول‌ها/بافت‌ها مقدار پروموترها به میزان قابل‌توجهی بیان شده باشند. تجزیه‌وتحلیل نتایج نشان می‌دهد که مقادیر بیان ژن با حد آستانه‌های مختلف مقادیر متفاوتی در ضریب سیلوئت ایجاد می‌کنند. برای این منظور نتایج با حدآستانه‌های متفاوتی بررسی شده است. همچنین برای تجزیه‌وتحلیل بهتر، نتایج حاصل شده با تحقیقات قبلی (خوشه‌بندی باینری و تجمعی) در معیارهای تعداد خوشه‌ها، سیلوئت و زمان اجرا نیز مقایسه شده است.

نتایج حاصل از اعمال روش پیشنهادی برای استخراج ارتباطات بین سلولی با مقدار حد آستانه بهینه $\theta = 0.2$ در جدول ۷ برای ۱۰۸ سلول در ۹ خوشه نشان داده شده است. این اطلاعات برای ارتباطات بین بافتی با مقدار حد آستانه بهینه $\theta = 0.1$ در جدول ۸ برای ۴۰ بافت در ۱۳ خوشه نشان داده‌شده است.

جدول ۷: مقایسه نتایج خوشه‌بندی با حد آستانه‌های مختلف برای بافت‌ها

روش‌ها	حد آستانه (θ)	۰/۰۵	۰/۱	۰/۱۵	**۰/۲**	۰/۲۵	۰/۳	۰/۳۵
خوشه‌بندی باینری	تعداد خوشه‌ها	۱۶	۱۹	۱۹	۱۲	۱۰	۱۲	۸
	شاخص سیلوئت	۰/۶۴۳	۰/۷۵۸	۰/۶۰۱	۰/۷۵۴	۰/۶۹۹	۰/۴۳۵	۰/۳۸۱
	زمان اجرا (ثانیه)	۶۷۱	۶۲۷	۶۰۱	۵۸۷	۵۷۷	۵۶۰	۵۴۷
خوشه‌بندی تجمعی	تعداد خوشه‌ها	۸	۸	۱۰	۹	۱۱	۶	۴
	شاخص سیلوئت	۰/۵۶۵	۰/۵۶۱	۰/۵۸۹	۰/۶۹۸	۰/۶۷۸	-۰/۷۰۱	-۰/۷۸۹
	زمان اجرا (ثانیه)	۲۰۶۷	۱۸۹۵	۱۸۰۶	۱۷۷۲	۱۷۱۵	۱۶۸۲	۱۶۷۵
خوشه‌بندی مبتنی بر گراف	تعداد خوشه‌ها	۸	۸	۸	۹	۹	۵	۲
	شاخص سیلوئت	۰/۵۹۲	۰/۵۹۲	۰/۶۱۸	**۰/۸۱۴**	۰/۷۴۸	-۰/۷۰۲	-۰/۹۵۶
	زمان اجرا (ثانیه)	۵۱۳	۵۰۶	۴۸۹	۴۶۸	۴۵۰	۴۴۱	۴۳۶

جدول ۸: مقایسه نتایج خوشه‌بندی با حد آستانه‌های مختلف برای بافت‌ها

روش‌ها	حد آستانه (θ)	۰/۰۵	**۰/۱**	۰/۱۵	۰/۲	۰/۲۵	۰/۳	۰/۳۵
خوشه‌بندی باینری	تعداد خوشه‌ها	۹	۸	۱۱	۱۰	۱۱	۹	۱۳
	شاخص سیلوئت	۰/۴۹۸	۰/۵۱۶	۰/۴۷۸	۰/۶۰۰	۰/۶۵۴	۰/۶۷۵	۰/۶۶۱
	زمان اجرا (ثانیه)	۴۶۸	۴۵۷	۴۴۱	۴۲۷	۴۲۰	۴۰۹	۳۹۸
خوشه‌بندی تجمعی	تعداد خوشه‌ها	۱۱	۹	۹	۱۵	۱۳	۱۳	۱۵
	شاخص سیلوئت	۰/۵۲۰	۰/۶۱۹	۰/۶۰۴	۰/۷۶۲	۰/۷۲۳	۰/۷۰۰	۰/۷۳۷
	زمان اجرا (ثانیه)	۱۶۸۱	۱۶۵۵	۱۶۱۲	۱۵۷۹	۱۵۶۲	۱۵۳۲	۱۵۰۶
خوشه‌بندی مبتنی بر گراف	تعداد خوشه‌ها	۱۳	۱۳	۱۱	۹	۴	۲	۱
	شاخص سیلوئت	۰/۷۰۶	**۰/۷۹۸**	۰/۷۱۸	۰/۶۴۴	-۰/۲۷۸	-۰/۵۹۷	-۰/۸۲۳
	زمان اجرا (ثانیه)	۴۰۶	۳۸۲	۳۷۰	۳۵۶	۳۴۹	۳۴۰	۳۳۱

با توجه به نتایج آزمایش، حد آستانه ۰/۲ دارای مقدار بهینه با ضریب سیلوئت ۰/۸۱۴ برای سلول‌ها می‌باشد. این در حالی است که در روش‌های خوشه‌بندی باینری و تجمعی به ترتیب بهترین ضریب سیلوئت ۰/۷۵۸ ۰/۶۹۸ است. نتایج خوشه‌بندی مبتنی بر گراف (روش پیشنهادی) با پارامتر آستانه بهینه در جدول ۹ نشان داده‌شده است. نماد C_i معرف سلول i-ام است که اسامی آن‌ها در پیوست ۱ آورده شده است.

با توجه به نتایج به‌دست‌آمده بالاترین میزان ارتباطات بین سلولی مربوط به سلول‌های hes3.gfp.embryonic.stem.cells ($C_{۸۲}$) و cd14.cd16..monocytes.2 ($C_{۲۹}$) با ۶۴۵۸۰ بیان ژن می‌باشد. در رتبه دوم و سوم به ترتیب سلول cd14..monocytes ($C_{۱۷}$) با ciliary.epithelial.cells ($C_{۴۹}$) و basophils ($C_{۱۵}$) با cd14..monocytes ($C_{۱۷}$) دارای بالاترین ارتباط بیان ژن می‌باشند.

این تعداد بیان ژن رفتار مشابه این دو سلول را در مواجه با بیماری‌های مختلف نشان می‌دهد. این اطلاعات می‌تواند در استخراج الگوهای رفتاری از یک ویروس خاص کمک کند. به‌طورکلی بیشتر ارتباطات در ژن ABLIM1 بیان‌شده و در ژن‌های TACC2 و KIAA1217 در رتبه‌های بعدی قرار دارند.

با توجه به نتایج آزمایش، حد آستانه ۰/۱ دارای مقدار بهینه با ضریب سیلوئت ۰/۷۸۹ برای بافت‌ها می‌باشد. این در حالی است که در روش‌های خوشه‌بندی باینری و تجمعی به ترتیب بهترین ضریب سیلوئت ۰/۶۷۵ و ۰/۷۶۲ است. نتایج خوشه‌بندی مبتنی بر گراف با پارامتر آستانه بهینه در جدول ۱۰ نشان داده‌شده است. نماد T_i معرف بافت i-ام مطابق پیوست ۲ می‌باشد. با توجه به نتایج به‌دست‌آمده بالاترین میزان ارتباطات بین بافتی مربوط به بافت‌های medulla.oblongata ($T_{۱۸}$) و testis ($T_{۳۳}$)

با ۷۵۳۵۶ بیان ژن می‌باشد. در رتبه دوم و سوم به ترتیب بافت locus.coeruleus ($T_{۱۴}$) با testis ($T_{۳۳}$) و medulla.oblongata ($T_{۱۸}$) با trachea ($T_{۳۹}$) دارای بالاترین ارتباط بیان ژن می‌باشند. به‌طورکلی بیشتر ارتباطات در ژن ACSL بیان‌شده و ژن‌های CELF و TCF7L در رتبه‌های بعدی قرار دارند.

جدول ۹: نتایج خوشه‌بندی با حدآستانه ۰/۲ برای سلول‌ها.

خوشه‌ها	نمونه‌ها (سلول‌ها)
خوشه ۱	$C_{۱۰}$
خوشه ۲	$C_{۱۱}$
خوشه ۳	$C_{۴۶}$
خوشه ۴	$C_{۴۵}$
خوشه ۵	$C_{۳۰},C_{۳۱},C_{۶۲}$
خوشه ۶	$C_{۳},C_{۹},C_{۳۴}$-$C_{۳۷},C_{۴۰},C_{۴۴}, C_{۸۰},C_{۸۴},C_{۸۵},C_{۹۰},C_{۹۸}$
خوشه ۷	$C_{۴},C_{۱۲}$-$C_{۱۵},C_{۱۷},C_{۱۹}$-$C_{۲۸}, C_{۳۳},C_{۳۸},C_{۳۹},C_{۴۱},C_{۴۲},C_{۴۹},C_{۵۲},C_{۵۸},C_{۷۸},$ $C_{۸۱},C_{۹۱},C_{۹۶},C_{۱۰۲},C_{۱۰۳}$
خوشه ۸	$C_{۱},C_{۶},C_{۴۷},C_{۶۳}$-$C_{۶۶},C_{۶۸},C_{۷۲},C_{۷۴}$-$C_{۷۶},C_{۷۹},C_{۹۴},C_{۹۹},C_{۱۰۰},C_{۱۰۷},C_{۱۰۸}$
خوشه ۹	$C_{۲},C_{۵},C_{۷},C_{۸},C_{۱۶},C_{۱۸},C_{۲۹},C_{۳۲},C_{۴۳},C_{۴۸},C_{۵۰},C_{۵۱},C_{۵۳}$-$C_{۵۷},C_{۵۹}$-$C_{۶۱},C_{۶۷},C_{۶۹}$- $C_{۷۱},C_{۷۳},C_{۷۷},C_{۸۲},C_{۸۳},C_{۸۶}$-$C_{۸۹},C_{۹۲},C_{۹۳},C_{۹۵},C_{۹۷},C_{۱۰۱},C_{۱۰۴}$-$C_{۱۰۶}$

جدول ۱۰ : نتایج خوشه‌بندی با حدآستانه ۰/۱ برای بافت‌ها.

خوشه‌ها	نمونه‌ها (بافت‌ها)
خوشه ۱	$T_{۳},T_{۱۵},T_{۳۲}$
خوشه ۲	$T_{۱۲},T_{۲۸},T_{۳۰},T_{۳۵},T_{۳۶},T_{۳۷},T_{۳۸},T_{۳۹}$
خوشه ۳	$T_{۴},T_{۲۰},T_{۲۳},T_{۲۴}$
خوشه ۴	$T_{۸},T_{۱۳},T_{۲۱},T_{۲۶},T_{۲۷},T_{۳۳},T_{۴۰}$
خوشه ۵	$T_{۲۲},T_{۲۵}$
خوشه ۶	$T_{۱۰}$
خوشه ۷	$T_{۱۱},T_{۱۶},T_{۱۷}$
خوشه ۸	$T_{۵}$
خوشه ۹	$T_{۱۴}$
خوشه ۱۰	$T_{۹},T_{۱۸},T_{۳۱},T_{۳۴}$
خوشه ۱۱	$T_{۱},T_{۷},T_{۱۹}$
خوشه ۱۲	$T_{۲۹}$
خوشه ۱۳	$T_{۲},T_{۶}$

۶- نتیجه‌گیری

ارتباطات بین سلول‌ها یا بافت‌ها به شناسایی بیماری‌های مختلف و عوامل آن‌ها کمک خواهد کرد. درواقع ارتباط بین سلول‌ها/بافت‌ها روابط وراثتی بین بیماران را نشان می‌دهد. این روابط به شناسایی نقاط مشترک بدن که تحت تأثیر بیماری‌های مختلفی قرار می‌گیرد، کمک می‌کند. در این تحقیق تشخیص ارتباط‌های بین سلولی/ بافتی در بیماری‌های مختلف با ترکیب نرمال‌سازی RLE، روش ویلکاکسون و الگوریتم خوشه‌بندی مبتنی بر گراف Louvain ارائه شده است. ارزیابی عملکرد الگوریتم خوشه‌بندی پیشنهادی با شاخص سیلوئت، دقت بالای آن را به اثبات رسانده است. برای به دست آوردن ارتباط بین سلولی/بافتی از روش ویلکاکسون و با pvalue بزرگ‌تر از ۰/۰۵ استفاده می شود. خروجی ماتریس همبستگی تعداد بیان ژن‌های مشترک بین هر دو نمونه سلول/بافت را نشان می‌دهد. این ماتریس ، یک ماتریس وزن‌دار است که به‌صورت یک گراف کامل بیان می‌شود. الگوریتم Louvain از یک روش حریصانه بر روی پیمانگی استفاده می‌کند و ارتباطات نهایی بین سلولی/بافتی را از طریق خوشه‌بندی استخراج می‌کند. روش پیشنهادی روی مجموعه داده فانتوم ۵ آزمایش شده است. نتایج نشان می‌دهد که به‌طور میانگین الگوریتم خوشه‌بندی پیشنهادی بر اساس شاخص سیلوئت به‌خوبی ارتباطات بین سلول ها/بافت‌ها را شناسایی می‌کند. از مزیت‌های روش پیشنهادی کاهش حجم داده‌های بیان ژن از طریق نگاشت آنها به صورت گراف است. این رویکرد باعث استفاده از الگوریتم‌های خوشه‌بندی پایه مباتی بر گراف بر روی داده‌های بیان ژن می شود. به‌جای محاسبه مشابهت بین اشیاء توسط یک حد آستانه، از چندین آستانه مختلف استفاده‌شده که درنهایت آ ستانه ۰/۲ برای سلول‌ها و آ ستانه ۰/۱ برای بافت‌ها بهترین نتایج را با توجه به شاخص سیلوئت حاصل نموده است. ارتباطات سلول به سلول نهایی در ۹ خوشه برای ۱۰۸ سلول و ارتباطات بافت به بافت نهایی در ۱۳ خوشه برای ۴۰ بافت، در بهترین حالت گزارش‌شده است.

پیوست ۱ : اسامی کامل سلول‌ها (C_i) در جدول ۹.

C۱: x293slam.rinderpest.infection,
C۲: arpe.19.emt.induced.with.tgf.beta.and.tnf.alpha,
C۳: adipocyte...breast,
C۴: adipocyte...omental,
C۵: adipocyte...subcutaneous,
C۶: adipocyte.differentiation,
C۷: alveolar.epithelial.cells,
C۸: amniotic.epithelial.cells,
C۹: anulus.pulposus.cell,
C۱۰: aortic.smooth.muscle.cell.response.to.fgf2,
C۱۱: aortic.smooth.muscle.cell.response.to.il1b,
C۱۲: astrocyte...cerebellum,
C۱۳: astrocyte...cerebral.cortex,
C۱۴: b.lymphoblastoid.cell.line..gm12878.encode,
C۱۵: basophils,
C۱۶: bronchial.epithelial.cell,
C۱۷: cd14..monocytes,
C۱۸: cd14..monocyte.derived.endothelial.progenitor.cells,
C۱۹: cd14..monocytes...mock.treated,
C۲۰: cd14..monocytes...treated.with.b.glucan,
C۲۱: cd14..monocytes...treated.with.bcg,
C۲۲: cd14..monocytes...treated.with.candida,
C۲۳: cd14..monocytes...treated.with.cryptococcus,
C۲۴: cd14..monocytes...treated.with.group.a.streptococci,
C۲۵: cd14..monocytes...treated.with.ifn...n.hexane,
C۲۶: cd14..monocytes...treated.with.salmonella,
C۲۷: cd14..monocytes...treated.with.trehalose.dimycolate..tdm.,
C۲۸: cd14..monocytes...treated.with.lipopolysaccharide,

C۲۹: cd14.cd16..monocytes,
C۳۰: cd14.cd16..monocytes.1,
C۳۱: cd14.cd16..monocytes.2,
C۳۲: cd19..b.cells..pluriselect.,
C۳۳: cd19..b.cells,
C۳۴: cd34.cells.differentiated.to.erythrocyte.lineage,
C۳۵: cd34..progenitors,
C۳۶: cd34..stem.cells...adult.bone.marrow.derived,
C۳۷: cd4..t.cells,
C۳۸: cd4.cd25.cd45ra..naive.regulatory.t.cells.expanded,
C۳۹: cd4.cd25.cd45ra..naive.regulatory.t.cells,
C۴۰: cd4.cd25.cd45ra..memory.regulatory.t.cells.expanded,
C۴۱: cd4.cd25.cd45ra..memory.conventional.t.cells.expanded,
C۴۲: cd4.cd25.cd45ra..memory.conventional.t.cells,
C۴۳: cd8..t.cells..Pluriselect.
C۴۴: cd8..t.cells,
C۴۵: cobl.a.rinderpest.infection,
C۴۶: cobl.a.rinderpest..c..infection,
C۴۷: cardiac.myocyte,
C۴۸: chondrocyte...de.diff,
C۴۹: ciliary.epithelial.cells,
C۵۰: corneal.epithelial.cells,
C۵۱: dendritic.cells...monocyte.immature.derived,
C۵۲: dendritic.cells...plasmacytoid,
C۵۳: endothelial.cells...aortic,
C۵۴: endothelial.cells...artery,
C۵۵: endothelial.cells...lymphatic,
C۵۶: endothelial.cells...microvascular,
C۵۷: endothelial.cells...thoracic,
C۵۸: endothelial.cells...umbilical.vein,
C۵۹: endothelial.cells...vein,
C۶۰: eosinophils,
C۶۱: esophageal.epithelial.cells,
C۶۲: fibroblast...aortic.adventitial.donor۲...cytoplasmic.fraction. C۶۳: fibroblast...aortic.adventitial,
C۶۴: fibroblast...cardiac,
C۶۵: fibroblast...choroid.plexus,
C۶۶: fibroblast...conjunctival,
C۶۷: fibroblast...dermal,
C۶۸: fibroblast...gingival,
C۶۹: fibroblast...lung,
C۷۰: fibroblast...lymphatic,
C۷۱: fibroblast...mammary,
C۷۲: fibroblast...periodontal.ligament,
C۷۳: fibroblast...villous.mesenchymal,
C۷۴: fibroblast...skin.dystrophia.myotonica,
C۷۵: fibroblast...skin.normal,
C۷۶: fibroblast...skin.spinal.muscular.atrophy,
C۷۷: fibroblast...skin,
C۷۸: gingival.epithelial.cells,
C۷۹: h1.embryonic.stem.cells.differentiation.to.cd34..Hsc,
C۸۰: h9.embryoid.body.cells,
C۸۱: h9.embryonic.stem.cells,
C۸۲: hes3.gfp.embryonic.stem.cells,
C۸۳: hair.follicle.dermal.papilla.cells,
C۸۴: hair.follicle.outer.root.sheath.cells,
C۸۵: hep.2.cells.mock.treated,
C۸۶: hep.2.cells.treated.with.streptococci.strain.5448,
C۸۷: hep.2.cells.treated.with.streptococci.strain.jrs4,
C۸۸: hepatic.sinusoidal.endothelial.cells,
C۸۹: hepatic.stellate.cells..lipocyte.,
C۹۰: hepatocyte,
C۹۱: k562.erythroblastic.leukemia.response.to.hemin,
C۹۲: keratinocyte...epidermal,
C۹۳: keratocytes,
C۹۴: lens.epithelial.cells,
C۹۵: lymphatic.endothelial.cells.response.to.vegfc,
C۹۶: mcf7.breast.cancer.cell.line.response.to.egf1,
C۹۷: mcf7.breast.cancer.cell.line.response.to.hrg,
C۹۸: macrophage...monocyte.derived,
C۹۹: mast.cell,
C۱۰۰: melanocyte...dark,
C۱۰۱: melanocyte...light,
C۱۰۲: melanocyte,
C۱۰۳: meningeal.cells,
C۱۰۴: mesenchymal.stem.cells...adipose,
C۱۰۵: mesenchymal.stem.cells...bone.marrow,
C۱۰۶: mesenchymal.stem.cells...umbilical,
C۱۰۷: mesothelial.cells,
C۱۰۸: monocyte.derived.macrophages.response.to.lps.

پیوست ۲: اسامی کامل بافت‌ها (T_i) در جدول ۱۰.

T_1: mesenchymal.stem.cells...adipose,
T_2: adipose,
T_3: amygdala,
T_4: brain,
T5: caudate.nucleus,
T_6: cerebellum,
T_7: colon,
T_8: duodenum,
T_9: globus.pallidus,
T_{10}: heart,
T_{11}: hippocampus,
T_{12}: kidney,
T_{13}: liver,
T_{14}: locus.coeruleus,
T_{15}: lung,
T_{16}: medial.frontal.gyrus,
T_{17}: medial.temporal.gyrus,
T_{18}: medulla.oblongata,
T_{19}: mesenchymal.precursor.cell...adipose,
T_{20}: occipital.cortex,
T_{21}: occipital.lobe,
T_{22}: parietal.lobe,
T_{23}: pineal.gland,

$T_{۲۴}$: pituitary.gland,

$T_{۲۵}$: putamen,

$T_{۲۶}$: skeletal.muscle,

$T_{۲۷}$: skin,

$T_{۲۸}$: small.intestine,

$T_{۲۹}$: spinal.cord,

$T_{۳۰}$: spleen,

$T_{۳۱}$: substantia.nigra,

$T_{۳۲}$: temporal.lobe,

$T_{۳۳}$: testis,

$T_{۳۴}$: thalamus,

$T_{۳۵}$: throat,

$T_{۳۶}$: thymus,

$T_{۳۷}$: thyroid,

$T_{۳۸}$: tongue,

$T_{۳۹}$: trachea,

$T_{۴۰}$: uterus.

مراجع

[1] C. Pizzuti and S. E. Rombo, *Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods*, Bioinformatics. 2014.

[2] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, "Mixing local and global information for community detection in large networks," in Journal of Computer and System Sciences, 2014, vol. 80, no. 1, pp. 72–87.

[3] J. Xie and B. K. Szymanski, *Towards linear time overlapping community detection in social networks*, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2012.

[4] N. Ozaki, H. Tezuka, and M. Inaba, *A Simple Acceleration Method for the Louvain Algorithm*, Int. J. Comput. Electr. Eng., 2016.

[5] K. Macropol, T. Can, and A. K. Singh, *RRW: Repeated random walks on genome-scale protein networks for local cluster discovery*, BMC Bioinformatics, 2009.

[6] T. Qin and K. Rohe, *Regularized spectral clustering under the degree-corrected stochastic blockmodel*, Adv. Neural Inf. Process. Syst., 2013.

[7] A. Lancichinetti, S. Fortunato, and J. Kertész, *Detecting the overlapping and hierarchical community structure in complex networks*, New J. Phys., 2009.

[8] S. Bahadori and P. Moradi, *A local Random Walk method for identifying communities in social networks*, in 7th Conference on Artificial Intelligence and Robotics, IRANOPEN 2017, 2017.

[9] G. J. McLachlan, R. W. Bean, and D. Peel, *A mixture model-based approach to the clustering of microarray expression data*., Bioinformatics, 2002.

[10] M. B. Gorzałczany, F. Rudziński, and J. Piekoszewski, *Gene expression data clustering using tree-like SOMs with evolving splitting-merging structures*, in Proceedings of the International Joint Conference on Neural Networks, 2016.

[11] A. Csikász-Nagy, *Computational systems biology of the cell cycle*, Briefings in Bioinformatics. 2009.

[12] J. Hofbauer and K. Sigmund, *Evolutionary game dynamics*, Bulletin of the American Mathematical Society. 2003.

[۱۳] مجید محمدپور و حمید پروین، «الگوریتم ژنتیک آشوب گونه مبتنی بر حافظه و خوشه بندی برای حل مسائل بهینه سازی پویا»، مجله مهندسی برق دانشگاه تبریز، دوره ۴۶، شماره ۳، صفحه ۲۹۹-۳۱۸، تبریز، پاییز ۱۳۹۵.

[۱۴] سمیرا رفیعی و پرهام مرادی، «بهبود عملکرد الگوریتم خوشه‌بندی فازی سی- مینز با وزن‌دهی اتوماتیک و محلی ویژگی‌ها»، مجله مهندسی برق دانشگاه تبریز، دوره ۴۶، شماره ۲، صفحه ۷۵-۸۶، تبریز، تابستان ۱۳۹۵.

[15] D. M. Lane and A. Sándor, *Designing Better Graphs by Including Distributional Information and Integrating Words, Numbers, and Images*, Psychol. Methods, 2009.

[16] D. Chen, Y., Kamath, G., Suh, C., & Tse, "Community recovery in graphs with locality," in In International Conference on Machine Learning, 2016, pp. 689–698.

[17] C. Chekuri and A. Goldberg, "Experimental study of minimum cut algorithms," in SODA '97 Proceedings of the eighth annual ACM-SIAM symposium on Discrete algorithms, 1997, pp. 324–333.

[18] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer series in statistics Springer, Berlin, 2001.

[19] R. Babers, A. E. Hassanien, and N. I. Ghali, "A nature-inspired metaheuristic Lion Optimization Algorithm for community detection," in 2015 11th International Computer Engineering Conference: Today Information Society What's Next?, ICENCO 2015, 2016.

[20] S. Fortunato, *Community detection in graphs*, Physics Reports. 2010.

[21] M. E. J. Newman, *Finding community structure in networks using the eigenvectors of matrices*, Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys., 2006.

[22] S. U. Rehman, A. U. Khan, and S. Fong, "Graph mining: A survey of graph mining techniques," in Seventh International Conference on Digital Information Management (ICDIM 2012), 2012.

[23] W. Maharani and A. A. Gozali, *Collaborative Social Network Analysis and Content-based Approach to Improve the Marketing Strategy of SMEs in Indonesia*, in Procedia Computer Science, 2015.

[24] K. H. Li, P., Piao, Y., Shon, H. S., & Ryu, "Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data," BMC Bioinformatics, vol. 16, no. 1, p. 347, 2015.

[25] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, *Mapping and quantifying mammalian transcriptomes by RNA-Seq*, Nat. Methods, 2008.

[26] C. Trapnell et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation*, Nat. Biotechnol., 2010.

[27] L. C. Gandolfo and T. P. Speed, *RLE plots: Visualizing unwanted variation in high dimensional data*, PLoS One, 2018.

[28] A. R. R. Forrest et al., *A promoter-level mammalian expression atlas*, Nature, 2014.

[29] F. Wilcoxon, *Individual Comparisons by Ranking Methods*, Biometrics Bull., 1945.

[30] S. Aranganayagi and K. Thangavel, "Clustering categorical data using Silhouette coefficient as a relocating measure," in Proceedings - International Conference on Computational Intelligence and Multimedia Applications, ICCIMA 2007, 2008, vol. 2, pp. 13–17.

[31] Y. Pan, *Inferring Mechanism-Based Gene Regulatory Network Models from Expression and Sequence Data*, 2009.

[32] E. Côme and P. Latouche, "Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood," Stat. Modelling, vol. 15, no. 6, pp. 564–589, 2015.

[33] J. Y. Jiang, R. J. Liou, and S. J. Lee, *A fuzzy self-constructing feature clustering algorithm for text classification*, IEEE Trans. Knowl. Data Eng., 2011.