دانشور
پزشکی

دوماهنامه علمی–پژوهشی
دانشگاه شاهد
سال هفدهم– شماره ۸۸
شهریور ۱۳۸۹

# مقایسه نتایج خوشه‌بندی سلسله مراتبی و غیرسلسله مراتبی پروتئین‌های مرتبط با سرطان‌های مری، معده و کلون براساس تشابهات تفسیر هستی‌شناسی ژنی

**نویسندگان: حمید علوی مجد[1*]، یلدا زرنگارنیا[2]، مصطفی رضایی طاویرانی[3]، نصیبه خیر[4]، علی‌اکبر خادم معبودی[5]**

۱. دانشیار گروه آمار زیستی دانشکده پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی،تهران،ایران

۲. کارشناس ارشد آمار زیستی، مرکز تحقیقات پروتئومیکس دانشگاه علوم پزشکی شهید بهشتی،تهران

۳ . دانشیار گروه علوم پایه دانشکده پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی،تهران،ایران

٤. کارشناس ارشد علوم سلولی ملکولی، مرکز تحقیقات پروتئومیکس دانشگاه علوم پزشکی شهید بهشتی،تهران ،ایران

۵. استادیارگروه آمار زیستی دانشکده پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی،تهران

*نویسنده مسئول: E-mail: alavimajd@gmail.com

وصول: ۸۹/۴/۱۶
آخرین اصلاحات: ۸۹/۸/۱
پذیرش: ۸۹/۸/۸

## چکیده

**مقدمه و هدف:** به‌کارگیری روش‌های پروتئومیک و پیشرفت پژوهش‌های با توان بالای پروتئین‌ها زمینه‌ساز نیاز به روش‌های تازه در تحلیل‌های بیوانفورماتیک نتایج آزمایشگاهی‌شده‌است. تحلیل خوشه‌ای، روش آماری مناسبی است که می‌تواند در تحلیل این‌گونه داده‌ها، استفاده‌شود. هدف از این پژوهش، ارزیابی کارایی روش خوشه‌بندی فازی در تحلیل پروتئین‌های مرتبط با سرطان‌های دستگاه گوارش بوده‌است.

**مواد و روش‌ها:** در این پژوهش، پروتئین‌های شناسایی‌شده مرتبط با سرطان‌های مری، معده و کلون با استفاده از روش‌های خوشه‌بندی سلسله مراتبی و غیرسلسله مراتبی تحلیل و بررسی‌شدند. براساس هر یک از ابعاد هستی‌شناسی ژنی، پروتئین‌ها را خوشه‌بندی کرده، نتایج را با یکدیگر مقایسه کردیم.

**یافته‌ها:** نتایج به‌دست‌آمده نشان داد، عملکرد دو روش خوشه‌بندی سلسله مراتبی و تقسیم‌بندی به یکدیگر نزدیک بوده‌است. با وجود چشمگیر نبودن عرض سایه‌نمای کل خوشه‌بندی‌ها، بیشتر پروتئین‌های در هر خوشه، دارای اشتراکات بیولوژیکی قابل توجهی هستند. نتایج به‌دست‌آمده نشان می‌دهد، روش‌های خوشه‌بندی توانسته‌اند الگوهای تفسیری جدیدی را که قبل از خوشه‌بندی پروتئین‌ها مشخص نشده‌بودند، آشکار سازند.

**نتیجه‌گیری:** با بررسی نتایج حاصل از خوشه‌بندی PAM و خوشه‌بندی سلسله مراتبی، مشخص شد، نتایج این دو خوشه‌بندی بسیار مشابه یکدیگر هستند. شاید بتوان گفت، روش PAM به دلیل معرفی نماینده‌ای برای هر خوشه، موشکافانه‌تر عمل کرده، در حالی‌که روش سلسله مراتبی ساده‌تر بوده‌است. همچنین به نظر می‌رسد، پروتئین‌هایی که براساس تشابهات جایگاه سلولی در کنار یکدیگر قرار می‌گیرند، دارای تشابهات بیولوژیکی و عملکردی نیز هستند که باید این مسئله بیشتر بررسی‌شود.

**واژگان کلیدی:** بیوانفورماتیک، تفسیر هستی‌شناسی ژنی (Gene Ontology)، خوشه‌بندی، سرطان دستگاه گوارش

## مقدمه

سرطان، یکی از عمده علل اختلالات، مرگ و میر و ناتوانی در سراسر جهان است. براساس آمار سازمان بهداشت جهانی تا سال ۲۰۲۰ میلادی، شیوع سرطان در موارد جدید با پنجاه درصد افزایش روبه‌رو خواهدشد و این بیماری در سال ۲۰۳۰ اولین عامل مرگ و میر در دنیا به شمار خواهد آمد (۱و۲). از جمله سرطان‌های شایع در جهان، سرطان‌های دستگاه گوارش هستند که در کشورهای در حال توسعه، از شایع‌ترین سرطان‌های حال حاضر به‌شمارمی‌روند. به‌طورمثال، سرطان معده اکنون به تنهایی نزدیک به ده درصد کل سرطان‌ها را در جهان تشکیل می‌دهد و یکی از شایع‌ترین انواع سرطان‌ها است (۳). بدیهی است، حجم وسیعی از مطالعات پزشکی معطوف به شناسایی عوامل مؤثر در ابتلا به سرطان و همچنین درمان آن و به‌طور کلی، یافتن روش‌هایی برای پیش‌گیری از ابتلا به انواع سرطان شده‌است. در این میان، شناسایی بیومارکرهای وابسته به بیماری که پیش از ظهور علائم بیماری خود را نشان دهند، اهمیت ویژه‌ای خواهدداشت.

پروتئین‌ها به دلایل مختلف، بیومارکرهای بسیار خوبی به‌شمارمی‌روند. با بررسی پروتئین‌ها می‌توان به طور مستقیم، عوامل مؤثر در بیماری را مطالعه کرد. از طرفی، تنها با مطالعه پروتئین‌ها می‌توان تغییرات پس از ترجمه را بررسی کرد (٤). با پیشرفت آزمایش‌های پروتئومیکی با توان بالا، مانند آرایه‌بندی پروتئین‌های تصفیه‌شده، لازم است پروتئین‌ها را به صورت جمعی مطالعه کنیم. البته این برخلاف روش سنتی تحلیل در هر بار یک پروتئین است. باید دانست، زمانی که تعداد پروتئین‌ها زیاد باشد، امکان این‌که در هر بار آزمایش یکی از آن‌ها را بررسی کنیم، وجود ندارد. علاوه بر این، شاید الگوهای جالبی در هر مجموعه از پروتئین‌ها وجودداشته‌باشد که اگر هر بار یکی از آن‌ها را تحلیل کنیم به نظر نیایند. ابزار در دسترس و بدون محدودیت بسیاری، برای تحلیل مجموعه‌های پروتئین‌ها وجود دارند اما اغلب آن‌ها نتایج حاصله را در یک تصویر ساده و شفاف و قابل تفسیر ارائه نمی‌دهند (۵).

بوریس آدریان و همکاران، در سال ۲۰۰٤ نرم‌افزار GO-Clust را معرفی کردند که از ساختار درختی پایگاه‌داده هستی‌شناسی ژنی به‌عنوان چارچوبی برای خوشه‌بندی داده‌های بیان ژن، استفاده می‌کند (٦). پوپسکو و همکاران، در سال ۲۰۰٤ به منظور یافتن عبارت‌هایی که بتوانند در خوشه‌بندی براساس تشابهات هستی‌شناسی ژنی نمایندگان خوبی برای خوشه‌ها باشند از روش خوشه‌بندی سلسله مراتبی و اندازه مشابهت های حاصل از BLAST و اندازه مشابهت فازی استفاده کردند (۷). هوگو و همکاران، در سال ۲۰۰٦ به منظور بررسی کارایی خوشه‌بندی پروتئین‌ها در تسریع مطالعه آن‌ها، پروتئین‌ها را به کمک اندازه مشابهت‌های حاصل از BLAST خوشه‌بندی کردند. آن‌ها پس از بررسی خوشه‌ها از لحاظ هستی‌شناسی ژنی، متوجه شدند مرکز هر خوشه می‌تواند اطلاعات پروتئین‌های خوشه را در برگیرد (۸). کریستین اواسکا و همکاران، در سال ۲۰۰۸ با اندازه‌گیری مشابهت‌های بین عبارات هستی‌شناسی ژنی و اجرای روش خوشه‌بندی سلسله مراتبی توانستند روشی ارائه‌دهند که در شناسایی سریع ژن‌هایی که دارای عبارات GO مشترک هستند، کمک کند (۹).

در این پژوهش، روش‌های خوشه‌بندی تقسیم‌بندی، حول میدوئیدها و خوشه‌بندی سلسله مراتبی تجمعی به منظور مطالعه پروتئین‌های مرتبط با سرطان‌های دستگاه گوارش، استفاده شده و هدف این بوده است تا بدانیم، آیا با خوشه‌بندی پروتئین‌ها براساس تشابهات بین تفاسیر هستی‌شناسی آنها می‌توان به الگوهایی در زیرمجموعه‌هایی از پروتئین‌ها دست یافت که در مطالعه جداگانه آن‌ها، به نظر نیامده باشند؟ همچنین با مقایسه نتایج خوشه‌بندی سلسله مراتبی و غیرسلسله مراتبی، عملکرد آن‌ها را در تفسیر بهتر پروتئین‌ها، با یکدیگر مقایسه کنیم.

## مواد و روش‌ها

در این پژوهش، از داده‌های جمع‌آوری‌شده در مرکز تحقیقات پروتئومیک دانشکده پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی، استفاده‌شده‌است. این داده‌ها، شامل حدود پانصد پروتئین شناسایی‌شده مرتبط با سرطان‌های مری، معده و کلون هستند که پس از بررسی‌های صورت‌گرفته تعداد هفده پروتئین، به‌عنوان پروتئین‌های دخیل مشترک در سرطان‌های مری، معده و کلون شناسایی‌شده‌اند (۱۰).

به منظور تحلیل این پروتئین‌ها براساس هستی‌شناسی ژنی، اطلاعات موردنیاز از وب‌سایت مربوط به اطلاعات هستی‌شناسی ژنی جمع‌آوری شد (۱۱). پروژه GO سه دسته واژگان را برای توصیف ژن و خواص محصول ژنی مانند پروتئین در هر ارگانیسم فراهم می‌کند. از لحاظ هستی‌شناسی، هر پروتئین را می‌توان از سه جهت بررسی کرد. این سه ویژگی عبارتند از: فرایند بیولوژیکی Biological Process (BP) ، جایگاه سلولی Cellular Component (CC) و عملکرد ملکولی Molecular Function (MF). براساس هر یک از این وجوه، گراف‌های GO از وب‌سایت Gene Ontology استخراج شدند.

برای اندازه‌گیری اندازه مشابهت‌های بین پروتئین‌ها، از میان روش‌های موجود، از روش simUI استفاده شد. براساس این روش، تشابه گرافیکی بین دو پروتئین عبارت است از: تعداد گره‌های مشترک در هر دو گراف GO تقسیم بر تعداد گره‌ها در اجتماع دو گراف با هم (۱۲و۱۳).

روش‌های خوشه‌بندی مشتمل بر روش‌های خوشه‌بندی سلسله مراتبی و غیرسلسله مراتبی هستند. روش‌های غیرسلسله مراتبی برای دسته‌بندی کردن اقلام به جای متغیرها به مجموعه‌ای از k خوشه طراحی شده است. یکی از روش‌های خوشه‌بندی غیرسلسله مراتبی، روش تقسیم‌بندی حول k میدوئید PAM (Partitioning Around Medoids) است. الگوریتم به‌کار رفته در این روش، براساس جست‌وجو برای یافتن K عنصر نماینده از بین مجموعه داده‌هاست که این k، عنصر نماینده را Medoids می‌نامند. در حقیقت، میدوئیدها نمایندگان خوشه‌ها هستند که باید بتوانند به خوبی ساختار داده‌ها را نشان دهند. نمایندگان هر خوشه کمترین مقدار متوسط عدم تشابه را با بقیه اعضای خوشه دارند. پس از مشخص شدن میدوئیدها k خوشه به این صورت ساخته می‌شود که هر عنصر به خوشه‌ای می‌رود که کمترین فاصله را با نماینده آن خوشه داشته باشد (۱۴و۱۵). پس از اجرای روش‌های خوشه‌بندی و حاصل شدن خوشه‌ها، می توان یک نمایش گرافیکی برای خوشه‌بندی به روش PAM داشت که آن‌را Silhouette یا سایه‌نما می‌نامیم. می‌توان به ازای هر خوشه، نمودار سایه‌نما داشت و با کنار هم قرار دادن آن‌ها کیفیت خوشه‌بندی‌ها را با یکدیگر مقایسه کرد. محور افقی این نمودار نشان‌دهنده عرض خوشه‌ها و محور عمودی، نحوه تعلق گرفتن عناصر به خوشه‌ها را نشان می‌دهد.

روش‌های خوشه‌بندی سلسله مراتبی با یک سری از تقسیم‌بندی‌های متوالی انجام می‌شود. روش‌های خوشه‌بندی سلسله مراتبی بر دو دسته‌اند: سلسله مراتبی تجمعی(Agglomerative) و سلسله مراتبی تقسیمی (Divisive). روشی را که ما استفاده کردیم، خوشه‌بندی سلسله مراتبی تجمعی ترکیب براساس پیوند متوسط (AGNES) است (۱۴و۱۵). در روش‌های سلسله مراتبی، نخست به تعداد داده‌ها، خوشه وجود دارد. آنگاه عناصری که بیشترین شباهت را دارند، دسته‌بندی شده و این دسته‌های اولیه براساس شباهت‌هایشان ظاهر می‌شوند. سرانجام، وقتی شباهت‌ها کاهش می‌یابد، تمام زیردسته‌ها به یک خوشه تبدیل می‌شوند. نتاج حاصل از این خوشه‌بندی را می‌توان در نمودار دندروگرام (Dendrogram) یا درختی نشان داد. در دندروگرام، محور عمودی فاصله بین خوشه‌ها را اندازه‌گیری می‌کند، ارتفاع هر یک از خوشه‌ها بیانگر آن است که، دو خوشه موردنظر در چه نقطه‌هایی با یکدیگر ترکیب شده‌اند (۱۴و۱۵).

یکی از مسائل مهم در خوشه‌بندی، تعیین تعداد بهینه خوشه‌هاست. مقدار کوچک k خوشه‌های بزرگی را نتیجه می‌دهد که ممکن است، روابطی را در آن خوشه نشان دهند که واقعاً وجودنداشته‌باشد. مقادیر بزرگ K نیز خوشه‌های کوچکی را نتیجه می‌دهد که ممکن است،

اطلاع مناسبی را در اختیار قرار ندهد زیرا روابط بین تعداد کمتری از اقلام را نشان می‌دهند. به منظور محاسبه تعداد بهینه خوشه‌ها، از روش Silcheck در Bioconductor براساس بیشینه کردن متوسط عرض سایه-نمای کل با تعیین تعداد حداکثر پنج خوشه، استفاده کرده‌ایم (۱۲). برای ارزیابی و اعتبارسنجی روش خوشه‌بندی PAM براساس تشابهات تفاسیر هستی-شناسی ژنی، نتایج خوشه‌بندی PAM را با یک روش تحلیلی مورد استفاده در تحلیل ریزآرایه DNA به کمک تعیین عبارت‌های GO غنی‌شده آماری مقایسه می‌کنیم. به‌این صورت که تفسیر GO یک زیر مجموعه از ملکول-ها با تفسیر GO مجموعه مرجع (UniProtKB) مقایسه می‌شود. اگر هر عبارت یا عبارت‌های نیایی آن (عبارت نیایی، والد عبارت موردنظر ما در گراف GO است.) بیشتر از حد معمول در هر زیرمجموعه نسبت به مجموعه مرجع رخ دهد، می‌گوییم آن عبارت GO غنی‌شده آماری است (۱۲و۱۶).

در روش خوشه‌بندی PAM، برای هر خوشه یک نماینده (میدوئید) تعیین می‌شود که برچسب هر خوشه تفسیر GO پروتئین میدوئید آن خوشه خواهد بود. اگر تفسیر GO پروتئین نماینده خوشه یک عبارت غنی‌شده آماری برای پروتئین‌های آن خوشه باشد، آنگاه می‌فهمیم برچسب پروتئین میدوئید به درستی تفسیر GO پروتئین‌ها را در آن خوشه نشان می‌دهد (۱۲).

برای آن‌که توانایی روش خوشه‌بندی PAM را در آشکارسازی الگوهای تفسیر جدید در مجموعه داده‌ها بررسی کنیم، عبارت‌های GO پروتئین‌های میدوئید را با عبارت‌های غنی‌شده آماری در کل مجموعه داده‌ها مقایسه می‌کنیم. چنانچه عبارات GO میدوئیدها عبارات غنی‌شده آماری در کل مجموعه داده‌ها باشند، آنگاه در حقیقت، به اطلاعات جدیدی در مجموعه داده‌ها دست نیافته‌ایم. اما اگر این عبارات در مجموعه کل داده‌های غنی‌شده آماری نباشند، آنگاه روش PAM الگوهای تفسیر جدیدی را در مجموعه داده‌هایمان آشکار کرده که قبلاً شناسایی نشده‌بودند (۱۲).

برای آن‌که توانایی روش خوشه‌بندی AGNES را در آشکارسازی الگوهای تفسیر جدید در مجموعه داده‌ها بررسی کنیم، عبارت‌های GO غنی‌شده آماری مربوط به پروتئین‌های هر خوشه را با عبارت‌های غنی‌شده آماری در کل مجموعه داده‌ها مقایسه می‌کنیم. چنانچه این عبارات در مجموعه کل داده‌ها به‌عنوان غنی‌شده آماری شناسایی نشده‌باشند، آنگاه روش AGNES الگوهای تفسیر جدیدی را در مجموعه داده‌ها آشکارساخته‌است.

به منظور بررسی درستی عملکرد روش خوشه‌بندی AGNESاز معیار همبستگی کوفنتیک (Cophenetic Correlation) که همبستگی بین فواصل اولیه بین پروتئین‌ها و فواصل کوفنتیک حاصل از خوشه‌بندی را محاسبه می‌کند، استفاده می‌کنیم (۱۷). برای مقایسه نتایج خوشه‌بندی‌های PAM و AGNES نیز از شاخص دان (Dunn Index)، ضریب همبستگی Hubertgamma و wb.ratio (average.within/average.between) استفاده می‌کنیم. (۱۸و۱۹و۲۰)

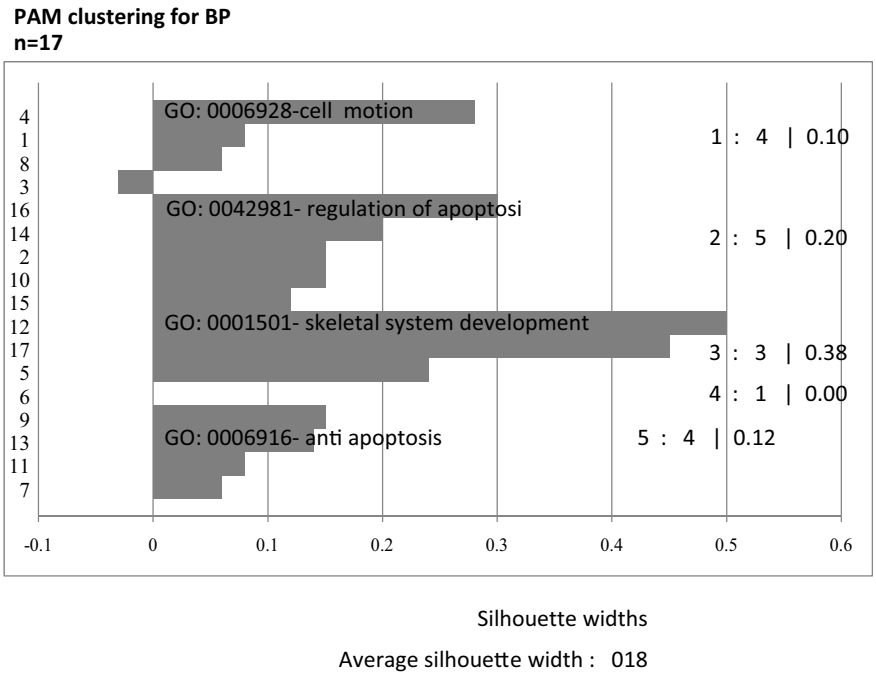
## نتایج

برای اجرای الگوریتم خوشه‌بندی از زبان برنامه-نویسی R استفاده شد (۲۱) و Packageهای موردنیاز نیز از وب‌سایت bioconductor دانلود شدند (۲۲). امتیازهای مشابهت براساس روش simUI برای تمام هفده پروتئین برای هر وجه GO به‌طور جداگانه محاسبه شد. تعداد بهینه خوشه‌ها براساس روش silcheck برای خوشه‌بندی به روش PAM به ترتیب ۵و ٤و ۵ خوشه براساس فرایند بیولوژیکی(BP) و مؤلفه سلولی(CC) و تابع ملکولی(MF) و برای خوشه‌بندی به روش AGNES برش دندروگرام در تعداد ٤ خوشه براساس فرایند بیولوژیکی(BP) و مؤلفه سلولی(CC) و تابع ملکولی(MF) تعیین شد.

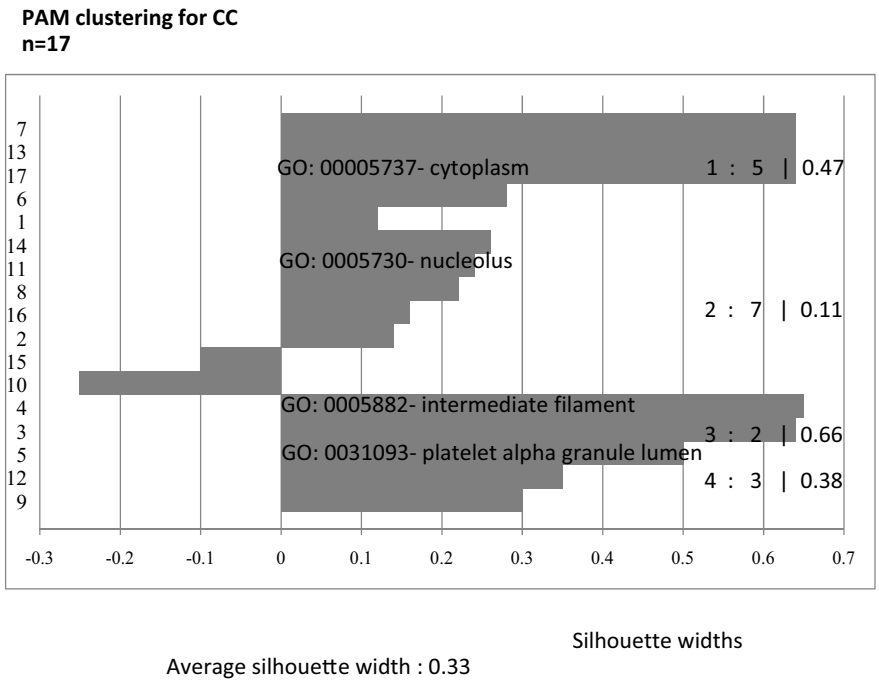
نتایج حاصل از خوشه‌بندی به روش PAM را می‌توان در جدول ۱ مشاهده کرد. نمودار سایه‌نمای این سه خوشه‌بندی در شکل‌های ۱، ۲ و ۳ ارائه‌شده‌اند. برچسب میدوئید هر خوشه عبارت GO مربوط به آن پروتئین است که در مقابل پروتئین میدوئید هر خوشه و. عرض سایه‌نمای کل در پایین هر خوشه‌بندی نوشته-شده‌است. عرض سایه‌نما مربوط به هر خوشه نیز در مقابل هر خوشه به همراه تعداد پروتئین‌ها متعلق به آن خوشه قرارگرفته‌است.

جدول ۱. نتایج خوشه‌بندی به روش PAM

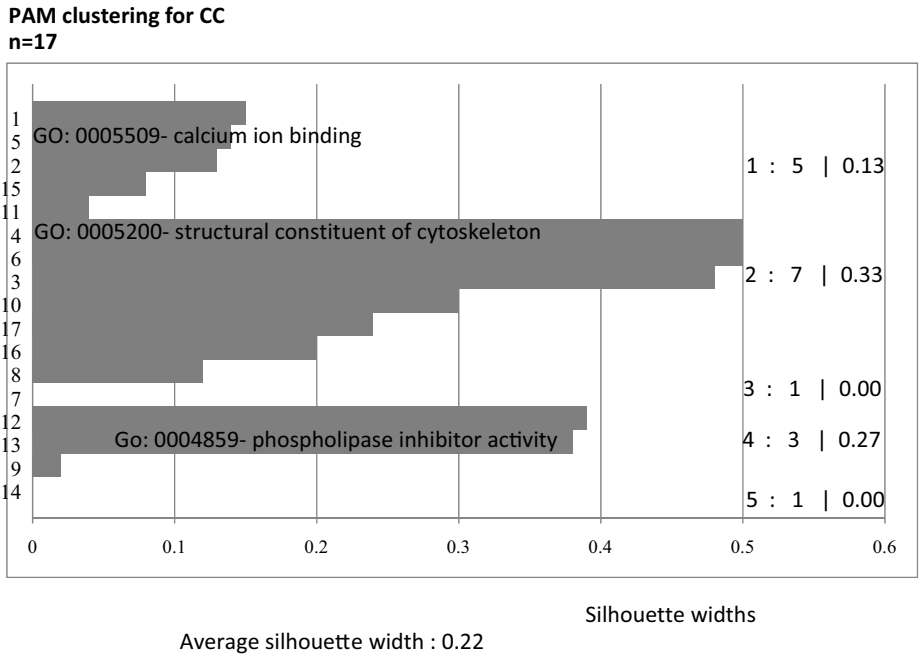
| Name | | BP | | CC | | MF | |
|---|---|---|---|---|---|---|---|
| cluster | Standard name | cluster | Silhouette width | cluster | Silhouette width | cluster | Silhouette width |
| ۱ | CAH2 | ۱ | ۰/۰۹۱ | ۱ | ۰/۱۲ | ۱ | ۰/۱۶ |
| ۲ | SODM | ۲ | ۰/۱۹ | ۲ | ۰/۱۴ | ۱ | ۰/۱۵ |
| ۳ | K2C8 | ۱ | -۰/۰۱۴۲ | ۳ | ۰/۶۶ | ۲ | ۰/۴۵ |
| ۴ | VIME | ۱ | ۰/۲۵ | ۳ | ۰/۶۶ | ۲ | ۰/۴۷ |
| ۵ | SPRC | ۳ | ۰/۲۲ | ۴ | ۰/۵۰ | ۱ | ۰/۱۵ |
| ۶ | DESM | ۴ | ۰/۰۰ | ۱ | ۰/۲۸ | ۲ | ۰/۴۷ |
| ۷ | PRDX2 | ۵ | ۰/۰۵ | ۱ | ۰/۶۵ | ۳ | ۰/۰۰ |
| ۸ | ACTB | ۱ | ۰/۰۶ | ۲ | ۰/۲۴ | ۲ | ۰/۱۴ |
| ۹ | AIAT | ۵ | ۰/۱۸ | ۴ | ۰/۳۰ | ۴ | ۰/۰۱۹ |
| ۱۰ | HSPBI | ۲ | ۰/۱۹ | ۲ | -۰/۲۴ | ۲ | ۰/۳۱ |
| ۱۱ | S10A9 | ۵ | ۰/۰۹۱ | ۲ | ۰/۲۵ | ۱ | ۰/۰۷ |
| ۱۲ | ANXA2 | ۳ | ۰/۴۸ | ۴ | ۰/۳۴ | ۴ | ۰/۴۰ |
| ۱۳ | ANXA5 | ۵ | ۰/۱۶ | ۱ | ۰/۶۵ | ۴ | ۰/۴۰ |
| ۱۴ | PCNA | ۲ | ۰/۲ | ۲ | ۰/۲۵ | ۵ | ۰/۰۰ |
| ۱۵ | CALR | ۲ | ۰/۱۷ | ۲ | -۰/۰۲۲ | ۱ | ۰/۰۹ |
| ۱۶ | PHB | ۲ | ۰/۲۷ | ۲ | ۰/۱۵ | ۲ | ۰/۲۰ |
| ۱۷ | TAGL | ۳ | ۰/۴۴ | ۱ | ۰/۶۵ | ۲ | ۰/۲۴ |
| Number of proteins clustered: | | ۱۷ | | ۱۷ | | ۱۷ | |



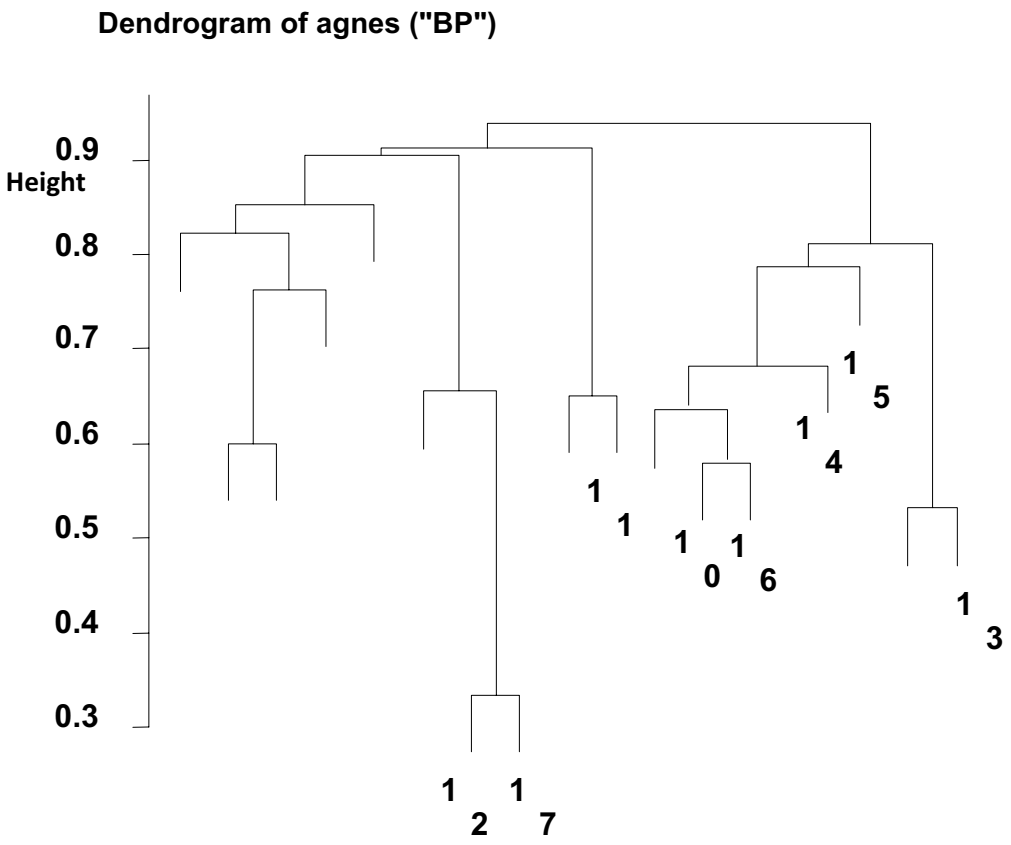

شکل ۱- نمودار سایه‌نما برای خوشه‌بندی PAM براساس BP



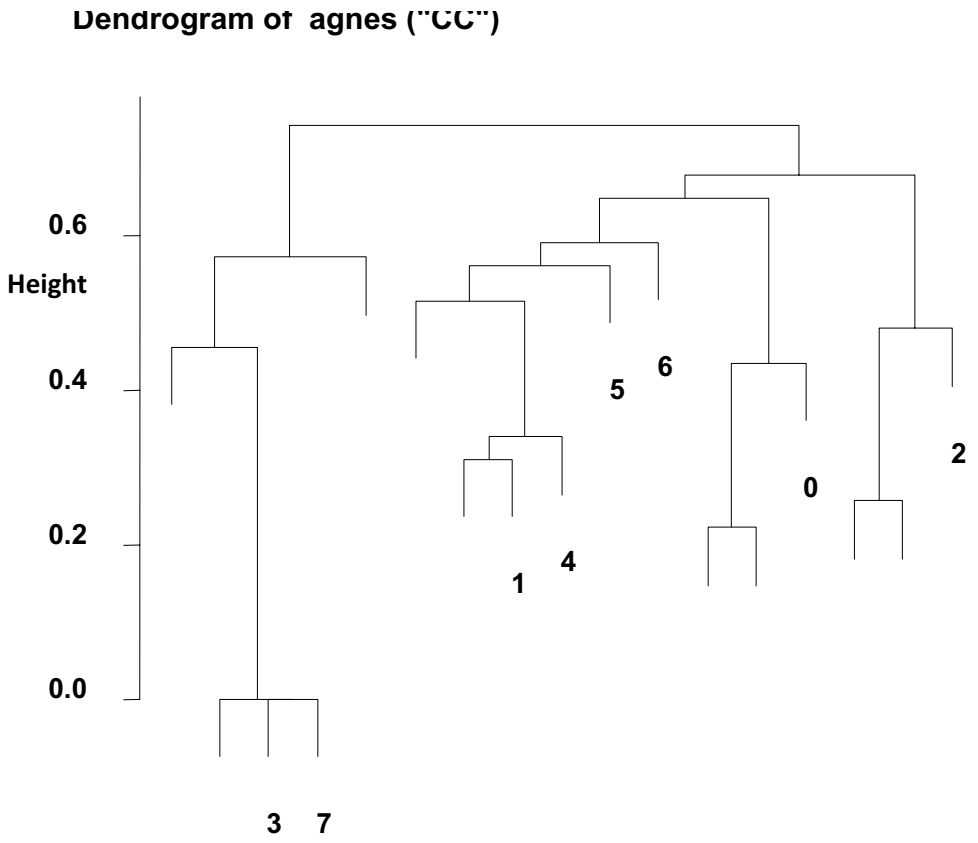

**شکل ۲- نمودار سایه‌نما برای خوشه‌بندی PAM براساس CC**



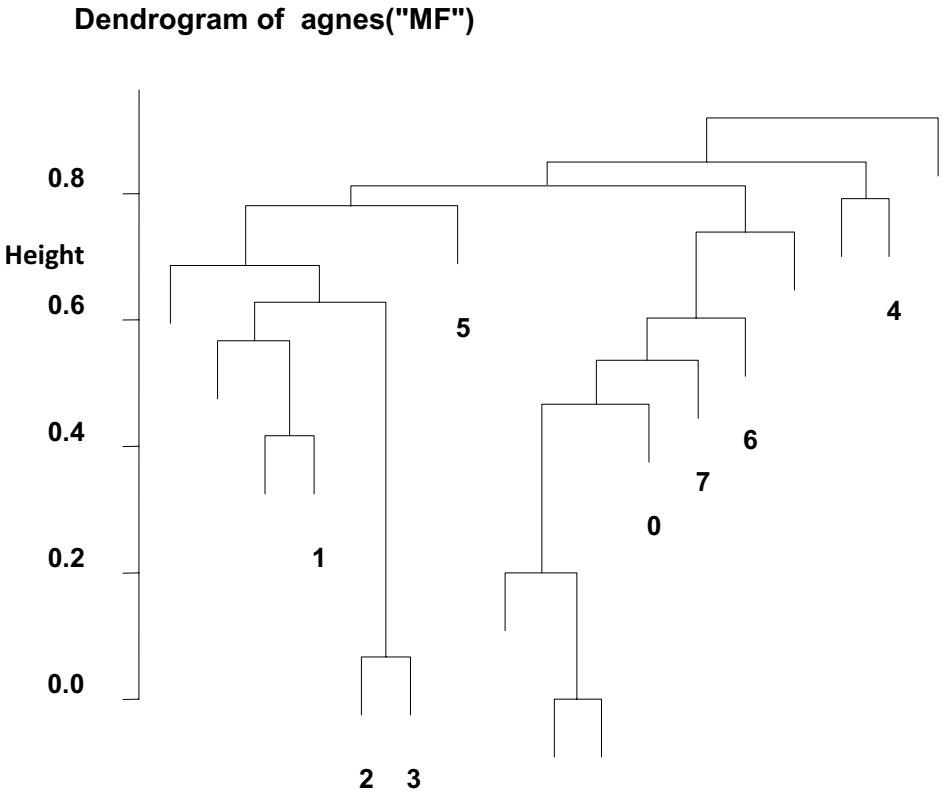

**شکل ۳- نمودار سایه‌نما برای خوشه‌بندی PAM براساس MF**




شکل ٤- نمودار دندروگرام برای خوشه‌بندی AGNES براساس BP




شکل ۵- نمودار دندروگرام برای خوشه‌بندی AGNES براساس CC




شکل ٦- نمودار دندروگرام برای خوشه‌بندی AGNES براساس MF

همان‌طور که دیده می شود ساختارهای خوشه‌بندی براساس MF و BP (BP $S_i^D$ =0.18,MF $S_i^D$ =0.22) (ساختارهای ضعیفی بوده‌اند و ساختار خوشه‌بندی براساس CC (CC $S_i^D$ =0.33) نیز ساختار به نسبت خوبی بوده‌است که البته به تحلیل‌های بیشتر نیاز دارد.

مقدار $S_i^D$ برای خوشه‌بندی بر اساس BPمقداری ضعیف و در کل تعداد ۳ مقدار از ٤ مقدار $S_i^C$مقادیر ضعیفی شده‌اند. خوشه سوم BPبا (0.38= $S_i^C$ ) خوشه به نسبت خوبی است. بررسی بیشتر نشان می‌دهد، هر سه این پروتئین‌ها در توسعه برخی از اندام‌ها نقش دارند. پروتئین‌های شماره ۵و ۱۲ در توسعه سیستم اسکلتی و پروتئین ۱۷ نیز در توسعه اندام ماهیچه‌ای نقش دارد.

علاوه بر این هر سه این پروتئین‌ها در اتصال به کلسیم نقش دارند. همچنین به‌رغم کوچک شدن عرض خوشه پنجم ( $S_i^C = 0.12$ )پروتئین های شماره ۹ و۱۱

و۱۳ درون این خوشه پروتئین‌های بازدارنده هستند. پروتئین‌های شماره ۹و ۱۱ در فرایند بیولوژیکی Inflammatory Response نقش دارند و پروتئین‌های شماره ۷و ۱۳ نیز تنظیم‌کننده‌های Apoptosis هستند. در خوشه اول نیز با وجود عرض خوشه کوچک ($S_i^C = 0.10$) پروتئین‌های شماره ٤ و۳ در اثرات متقابل بین ارگانیسم‌ها و پروتئین‌های شماره ٤و ۸ نیز در حرکت سلول نقش دارند. همچنین پروتئین های شماره ۱و ۳و ۸ هر سه در سیتوپلاسم موجودند و دو پروتئین شماره ۳و ٤ نیز در سلول‌های ماهیچه‌ای فعالیت می‌کنند. خوشه دوم نیز مقدار عرض خوشه کوچکی ($S_i^C = 0.20$) دارد، اما چهار پروتئین شماره ۲، ۱۰، ۱۵ و ١٦ موجود در این خوشه، در تنظیم Apoptosis نقش دارند.

مقدار $S_i^D$ برای خوشه‌بندی براساس MF نیز مقداری کمتر از ۰/۲۵ شده‌است که نشان می‌دهد، ساختار خوشه‌بندی چشمگیری به‌دست‌نیامده‌است. مقدار $S_i^C$ در خوشه‌های سوم و پنجم، صفر شده‌است زیرا در هر کدام از این خوشه‌ها تنها یک پروتئین قرارگرفته-است. در خوشه اول، $S_i^C = 0.13$ شده‌است که مقدار کوچکی است. به نظر می‌رسد، این خوشه از بقیه خوشه‌ها به خوبی جدانشده‌است. اما با وجود کوچکی عرض خوشه، بررسی نشان می‌دهد، پروتئین‌های شماره ۵ ، ۱۱و ۱۵ موجود در این خوشه، متصل‌شوندگان به یون‌های کلسیم و پروتئین‌های شماره ۱و ۱۵ نیز متصل‌شوندگان به یون‌های روی هستند، بنابراین به درستی در کنار یکدیگر قرارگرفته‌اند. مقادیر عرض خوشه در خوشه‌های دوم ($S_i^C$=0.33) و چهارم ($S_i^C$=0.27) از ۰/۲۵ بیشتر شده‌اند. در خوشه دوم، سه پروتئین شماره ۳، ٤ و ٦ در سلول‌های ماهیچه‌ای فعالیت می‌کنند و هر سه به خانواده Inter Mediate Filament تعلق‌دارند. در خوشه چهارم نیز پروتئین‌های شماره ۱۲و۱۳ جزء خانواده Annexin هستند. درضمن هر سه پروتئین‌های این خوشه پروتئین‌های بازدارنده آنزیمی هستند.

خوشه‌بندی براساس CCبا $S_i^C = 0.33$، ساختار خوشه‌بندی به نسبت متعادلی بوده است. خوشه‌های اول ($S_i^C = 0.47$)و سوم ($S_i^C = 0.66$) دارای عرض خوشه‌های خوبی هستند و به خوبی از بقیه خوشه‌ها جدا شده‌اند. هر پنج پروتئین موجود در این خوشه، در سیتوپلاسم فعالیت دارند. هر دو پروتئین خوشه سوم نیز در سیتوپلاسم و در قسمت Inter Mediate Filament فعالیت می‌کنند، بنابراین به درستی در کنار یکدیگر قرارگرفته‌اند. اما در مورد پروتئین‌های خوشه چهارم باید گفت، با وجود عرض خوشه متوسط ۰/۳۸ ، هر سه پروتئین شماره ۵ و۹ و۱۲ در فضای خارج از سلول و در Basement Membrance فعالیت دارند و به درستی در یک خوشه قرارگرفته‌اند. خوشه دوم نسبت به بقیه خوشه‌ها براساس تشابهات CC دارایعرض خوشه کمتری($S_i^C = 0.11$) است. با وجود کوچکی عرض خوشه، علاوه بر آن‌که پروتئین‌های شماره ۲، ۸، ۱۰، ۱۱و ۱۵ در سیتوپلاسم فعال هستند، پروتئین‌های شماره ۸، ۱۰، ۱۱، ١٤و ۱۵ نیز در نوکلئوز فعالیت دارند.

بررسی خوشه‌های حاصل از اجرای روش PAM ، براساس تشابهات CC نشان می‌دهد، عناصر موجود در این خوشه‌ها از لحاظ کارکرد و شرکت در فرایندهای بیولوژیکی بسیار به یکدیگر نزدیک هستند. پروتئین‌های شماره ۱۳و ۱۷ درون خوشه اول متصل‌شوندگان به یون‌های کلسیم و پروتئین شماره ۱ متصل شونده به یون روی است. اما خوشه دوم، عرض سایه‌نمای ($S_i^C = 0.11$) کوچکی دارد. با این حال، چهار پروتئین از پروتئین‌های این خوشه به شماره‌های ۲، ۱۰ ، ۱۵و ١٦ در تنظیم Apoptosis نقش دارند. خوشه سوم، عرض خوشه ($S_i^C = 0.66$) خوبی داشته‌است و هر دو پروتئین شماره ۳و ٤ موجود دراین خوشه در فرایند بیولوژیکی اثرات متقابل بین اندام‌ها نقش داشته، در سلول‌های ماهیچه‌ای فعالیت دارند. در خوشه چهارم نیز

پروتئین‌های شماره ۹و ۱۲ بازدارنده‌های آنزیمی هستند و پروتئین‌های شماره ۵ و ۱۲ در گسترش سیستم اسکلتی نقش دارند. در ضمن، پروتئین شماره ۹ متصل-شونده به پروتئاز و پروتئین‌های شماره ۵ و ۱۲ متصل‌شوندگان به یون‌های کلسیم هستند.

در مجموع می‌توان گفت با وجود آن‌که با استفاده از روش PAM خوشه‌هایی حاصل شده‌است که دارای عرض سایه‌نمای کل کوچکی هستند و ساختارهای خوشه‌بندی چشمگیری مشاهده نشده، اما خوشه‌های حاصل از لحاظ بیولوژیکی مفهوم داشته و به بررسی‌های بیشتر نیاز دارند. به‌خصوص در مورد خوشه‌بندی براساس مؤلفه سلولی، خوشه‌هایی حاصل شده‌است که عناصر آن‌ها از لحاظ کارکرد و شرکت در فرایندهای بیولوژیکی بسیار به یکدیگر نزدیک هستند.

برای ارزیابی اعتبار خوشه‌بندی‌های به‌دست‌آمده از روش PAM ، عبارات غنی‌شده آماری خوشه‌ها نرم‌افزار Term Enrichment موجود در وب‌سایت هستی‌شناسی ژنی (P- value=0/05) مشخص شدند. از کل تعداد چهارده خوشه ایجادشده به روش PAM، تعداد هشت خوشه دارای حداقل یک عبارات غنی‌شده آماری بودند. پس از مقایسه آن‌ها با عبارت‌های GO میدوئیدهای خوشه‌ها معلوم شد، برچسب میدوئیدهای صددرصد (هشت خوشه از هشت خوشه) از خوشه‌هایتولیدشده به روش PAM ، عبارت‌های غنی‌شده برای آن خوشه‌ها هستند. این، نشان‌دهنده مناسبت انتخاب عبارت‌های GO پروتئین‌های میدوئید به‌عنوان بر چسب خوشه‌هاست.

برای بررسی توانایی روش خوشه‌بندی PAM در شناسایی الگوهای جدید، در بررسی مجموعه داده‌ها پیش از خوشه‌بندی آن‌ها، Term Enrichment توانست تنها ۰/۲۹ (چهار خوشه از چهارده خوشه) از برچسب خوشه‌ها را به‌عنوان غنی‌شده آماری شناسایی کند، بنابراین روش PAM در خوشه‌بندی پروتئین‌ها براساس مشابهت‌های تفاسیر هستیشناسی ژنی و انتخاب عبارات GO پروتئین‌های میدوئید به‌عنوان برچسب خوشه‌ها، توانسته تفسیر بهتری از داده‌ها را در اختیار ما بگذارد و به‌عنوان ابزاری مفید در چهار خوشه (خوشه سوم مربوط به BP و خوشه اول، سوم و پنجم مربوط به MF) الگوهای تفسیر جدیدی را که قبلاً مشخص نشده‌اند، شناسایی کند.

برای اجرای روش خوشه‌بندی AGNES، ماتریس‌های مشابهت برای هر یک از وجوه GO به کمک روش simUI محاسبه و براساس هر یک از وجوه GO خوشه‌بندی اجرا شد. نمودارهای دندروگرام مربوط به این خوشه‌بندی را می‌توان در شکل‌های ٤ ،٥ و٦ مشاهده کرد. مقادیر ضریب ادغام (Agglomerative Coefficient) برای خوشه‌بندی براساس BP ،MF وCC به ترتیب ۰/۳٤، ۰/٤۹ و ۰/٥٦ به‌دست آمده‌اند. براساس مقادیر این ضریب می‌توان گفت، ساختار خوشه‌بندی براساس BP متعادل و براساس MF وCC خوب بوده-است. متوسط عرض نماهای حاصله براساس برش بین دو الی پنج خوشه نیز در جدول۲ نشان‌داده‌شده‌است. براساس مقادیر این جدول، تعداد بهینه چهار خوشه برای برش دندروگرام‌ها استفاده می‌شود.

جدول۲. متوسط عرض سایه‌نمای کل در خوشه‌بندی به روش AGNES

| Number of clusters | BP | MF | CC |
|---|---|---|---|
| 2 | 0/14 | 0/18 | 0/27 |
| 3 | 0/15 | 0/14 | 0/29 |
| 4 | 0/20 | 0/23 | 0/33 |
| 5 | 0/19 | 0/21 | 0/32 |

ساختار خوشه‌های به‌دست‌آمده از روش سلسله مراتبی، مانند نتایج به‌دست‌آمده از روش PAM، ساختار-هایی قوی نبوده‌اند. این ساختارها براساس MFو BP ($S_i^D$ =0/19,MF $S_i^D$ =0/21) (BP ضعیف و ساختار

خوشه‌بندی براساس CC ($S_i^D$ =0/32) CC) نیز به‌نسبت متوسط بوده‌است. نتایج به‌دست‌آمده با برش دندروگرام-ها در تعداد چهار خوشه، بسیار مشابه نتایج به‌دست‌آمده از روش PAMهستند. برای بررسی میزان تطابق نتایج خوشه‌بندی سلسله مراتبی و خوشه‌بندی PAM، اندازه مشابهت کسینوسی (Cosine Similarity Mmeasure)را اندازه‌گیری کردیم. با توجه به نتایج به‌دست‌آمده در جدول ۳، می‌توان گفت، نتایج برای خوشه‌بندی براساس CC بیشترین تطابق را داشته‌اند. به‌طور کل، نتایج بسیار به یکدیگر نزدیک بوده‌اند. از لحاظ بیولوژیکی، پروتئین‌هایی که در یک خوشه قرار گرفته‌اند مانند نتایج روش PAM، دارای اشتراکات فراوانی هستند و به درستی در کنار یکدیگر قرارگرفته‌اند.

جدول۳. میزان تطابق نتایج خوشه‌بندی PAM و AGNE

| | BP | MF | CC |
|---|---|---|---|
| PAM-AGNES | 0/71 | 0/84 | 0/88 |

جدول ٤. ارزیابی نتایج خوشه‌بندی PAM و AGNES

| GO aspects | Clustering methods | Dunn | wb ratio | Hubert gamma |
|---|---|---|---|---|
| BP | PAM | 0/76 | 0/77 | 0/73 |
| | AGNES | 0/76 | 0/80 | 0/74 |
| MF | PAM | 0/68 | 0/70 | 0/66 |
| | AGNES | 0/85 | 0/70 | 0/69 |
| CC | PAM | 0/61 | 0/65 | 0/65 |
| | AGNES | 0/63 | 0/62 | 0/67 |

براساس تفاسیر GO، به‌رغم ایجاد عرض سایه‌نمای کل ضعیف، به خوبی توانسته‌اند، الگوهای تفسیر جدیدی را شناسایی کنند که متفاوت از عبارت‌های بیش نشان-داده‌شده آماری در مجموعه کل داده‌ها بوده‌اند. بنابراین می‌توان گفت، اگر چه براساس منابع موجود در استفاده از روش PAM و AGNES خوشه‌های به‌وجودآمده با $S_i^C \le 0.25$ چندان قابل تفسیر نیستند اما چنین چیزی در مورد خوشه‌بندی پروتئین‌ها درست نیست و خوشه-های با مقادیر کم $S_i^C$ نیز از لحاظ بیولوژیکی خوشه-هایی قابل تفسیر هستند.

## بحث و نتیجه‌گیری

در مطالعه مجموعه‌ای از پروتئین‌ها، همواره هدف این است تا به الگوها و روابط و تفاسیری از آن‌ها دست یابیم که شاید ناشناخته مانده‌باشند. شریل ولتینگ و همکاران در سال ۲۰۰٦ مجموعه‌ای از داده‌های

از تعداد دوازده خوشه تولیدشده به روش AGNES، تعداد هفت خوشه دارای عبارت‌های غنی‌شده آماری بوده‌اند. از تعداد ۲۹ عبارت غنی‌شده آماری یافت‌شده در خوشه‌ها، تعداد ۲۱ عبارت در مجموعه کل داده‌ها به‌عنوان غنی‌شده آماری شناسایی نشده‌اند. بنابراین روش AGNES نیز به خوبی روش PAM توانسته، الگوهای جدیدی را شناسایی کند. به منظور مقایسه و ارزیابی دو روش PAM و AGNES از شاخص هایی همچون Dunn INDEX، نسبت wb و همبستگی Hubert Gamma استفاده شد. مقدار مطلوب برای این سه شاخص به ترتیب مقادیر دورتر از صفر، نزدیک به صفر و نزدیک به یک خواهد بود. با توجه به نتایج حاصل در جدول ٤، به دلیل نزدیکی مقادیر به یکدیگر هر دو روش، عملکرد بسیار نزدیک به یکدیگر داشته‌اند و می‌توان گفت، براساس این سه شاخص، ساختار خوشه‌بندی‌های حاصل چشمگیر نبوده‌اند. اما باید در نظر داشت، روش‌های خوشه‌بندی PAM و AGNES

نشده‌اند، پس لازم است، این خوشه بیشتر مرد تحلیل شود.

همچنین برچسب خوشه اول حاصل از اجرای روش PAM براساس MF نیز در مطالعه کلی داده‌ها به‌عنوان یک عبارت غنی‌شده آماری شناسایی نشد که این خوشه نیز باید بیشتر بررسی شود. در روش خوشه‌بندی AGNES نیز ۲۱ عبارت از ۲۹ عبارات GO غنی‌شده آماری یافت‌شده در خوشه‌ها، در مجموعه کل داده‌ها شناسایی نشده‌اند که به مطالعه بیشتر نیاز دارند. روش های خوشه بندی PAM و AGNES دارای نتایج بسیار مشابهی شدند و ارزیابی‌ها نیز نشان داد، هر دو روش عملکرد به‌نسبت مشابهی داشته‌اند. شاید بتوان گفت، روش PAM به دلیل معرفی نماینده‌ای برای هر خوشه دقیق‌تر عمل کرده در حالی‌که روش AGNES ساده‌تر بوده‌است. همچنین پروتئین‌های شماره ۳ و ٤ در تمامی خوشه‌بندی‌ها در کنار یکدیگر قرار گرفتند و به نظر می‌رسد، ارتباط خاصی بین این سرطان‌ها و این دو پروتئین برقرار باشد و لازم است این دو پروتئین بیشتر مطالعه شوند. بررسی بیشتر خوشه‌های حاصل از خوشه‌بندی‌های PAM و AGNES براساس CC، نشان داد، پروتئین‌های قرارگرفته در هر خوشه از لحاظ عملکرد و فرایندهای بیولوژیکی که در آن‌ها نقش دارند دارای تشابهات بسیاری هستند که خود تأمل‌برانگیز است. شاید با مطالعات بیشتر بتوان گفت، چنانچه مجموعه‌ای از پروتئین‌ها را داشته باشیم که مکان آن‌ها در سلول مشخص باشد، با خوشه‌بندی آن‌ها براساس CC، بتوان خوشه‌هایی تولید کرد که پیش‌بینی کنند، پروتئین‌های موجود در هر خوشه دارای کارکردهای مشترکی هستند یا در فرایندهای بیولوژیکی مشابهی نقش دارند.

## تشکر و قدردانی

این مقاله، حاصل طرح تحقیقاتی مشترک بین گروه آمار زیستی و مرکز تحقیقات پروتئومیکس دانشکده پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی است که به‌دین‌وسیله از هیئت رئیسه محترم دانشکده پیراپزشکی برای تأمین بودجه طرح پژوهشی، صمیمانه سپاسگزاری می‌شود.

پروتئینی مربوط به مخمر را به کمک روش خوشه‌بندی تقسیم بندی حول نماینده‌ها براساس تشابهات هستی-شناسی ژنی خوشه‌بندی کردند و داخل خوشه‌ها به الگوهای تفسیر جدید دست یافتند (۱۲). آن‌ها علاقه‌مند بودند، از روش PAM در خوشه‌بندی پروتئین‌های موجودات عالی‌تر استفاده کنند. در این پژوهش با اجرای روش خوشه بندی PAM برای تحلیل پروتئین-های مرتبط با سرطان‌های دستگاه گوارش مشخص شد، تحلیل خوشه‌ای توانسته، الگوهای تفسیر جدیدی را در مجموعه داده‌ها آشکار سازد. همچنین به منظور مقایسه روش خوشه‌بندی سلسله مراتبی با روش تقسیم‌بندی حول نماینده‌ها با اجرای هر دو روش و تعیین عبارت‌های غنی‌شده آماری مشخص شد، روش‌های خوشه‌بندی PAM و AGNES توانسته‌اند زیرمجموعه-های کوچکی را داخل داده‌ها ایجاد کنند که با تعیین عبارت GO پروتئین میدوئید به‌عنوان برچسب نماینده آن خوشه، به الگوهای جدیدی دست یابیم که با نگاه کلی به داده‌ها به چشم نیامده‌اند. این الگوهای شناسایی‌شده، ما را به مطالعات و پی‌گیری آزمایش‌های بیشتر برای بررسی مکانیسم‌ها و تأثیر وجود اثرات متقابل در غربالگری آرایه پروتئین‌ها ترغیب می‌کند.

از آنجا که در بررسی کل داده‌ها تنها توانستیم چهار عبارت از هشت عبارت برچسب میدوئید غنی‌شده آماری به‌دست‌آمده از خوشه‌بندی PAM را مشخص کنیم، خود نشان‌دهنده آن است که مطالعه کلی پروتئین‌ها نتوانسته در مورد ارتباطات بین زیرمجموعه-های کوچک‌تر، ایده‌های چندانی بدهد، بنابراین در چهار خوشه از هشت خوشه تولیدشده به روش PAM ، به نتایجی مبنی بر وجود الگوهای جدید کشف‌نشده در مجموعه داده‌ها می‌رسیم. با توجه به این که برچسب خوشه سوم حاصل از اجرای روش PAM ، براساس BP در بررسی کلی داده‌ها در وب‌سایت GO به‌عنوان عبارات غنی‌شده آماری شناسایی نشده‌است، بنابراین یک مجموعه از داده‌ها را شناسایی کرده‌ایم که نشان-دهنده وجود الگوهای جدیدی هستند که قبلاً مشاهده

## منابع


1- Tavoli A, Montazeri Ali, Mohagheghi MA, Roushan Rasoul, Tavoli Z, Meliani M. Role of information of cancer diagnosis in the quality of patients life are affected by digestive system cancer. Payesh, 6'th Year, No.3, 1386.

2- Are the number of cancer cases increasing or decreasing in the world?, April 2008, Available from World health organization: http://www.who.int/features/qa/15/en/index.html.

3- Parkin DM. Epidemiology of cancer: global patterns and Trends. Toxicol Lett 1998; 227:102-103.

4- M. Rezaei Tavirani, A. Marashi, F. Ghalanbar, M. Mostafavi. Preoteomics. Andishe Zohoor Publication, 1384.

5- Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 2005; 21:3587–3595.

6- Boris Andryan, Reinhard Schuch. Gene-Ontology-based clustering of gene expression data. Bioinformatics 2004; vol.20.no.16.

7- Popescu M, Keller IM, Mitchell JA, Bezdek JC. Functional summarization of gene produced clusters using Gene Ontology similarity measures. Intelligent Sensors, Sensor Networks and Information Processing Conference; 2004, 553-558.

8- Hugo Bastos, Daniel Faria, Catia pesquita and Andreo Flacao. Using GO terms to evaluate protein clustering. In ISMB/ECCB 2007 SIG Meeting Program Materials, Julay 2007; pages 107—110.

9- Kristian Ovaska, Marko Laakso and Sampsa Hautaniemi. Fast Gene Ontology based clustering for microarray experiments. BioData Mining 2008; 1:11.

10- Nasibeh Khaier, Mostafa Rezaei-Tavirani, Amir Rostami. Proteomics analysis of included proteins in esophagus, stomach and colon cancer.10th Iranian congress of Biochemistry and 3th international congress of Biochemistry and Molecular Biology; Tehran, 2009.

11- Search the Gene Ontology database, 2009 May-June, Available from http://amigo.geneontology.org/cgi-bin/amigo/go.cgi

12- Cheryl Wolting, C Jane McGlade, and David Trichler. Cluster analysis of protein array results via similarity of Gene Ontology Annotation. BMC Bioinformatics July 2006; Vol. 7.

13- Lord PW, Stevens RD, Brass A, Goble CA. investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 2003; 19:1275–1283.

14- Leonard Kaufman, Peter J. Rousseeuw. Finding Groups in data– An Introduction to Cluster Analysis. John Wiley & Sons, Inc. Publication, ISBN 0-471-73578-7.

15- Richard A. Johnson, Dean W. Wichern. Applied Multivariate Statistical Analysis. Prentice Hall, Englewood Cliffs, New Jersey;1988.

16- Susmita Datta, Somnath Datta. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, USA,2006.

17- Sneath, P.H.A. and Sokal, R.R. Numerical Taxonomy: The Principles and Practice of Numerical Classification, p. 278 ff; Freeman, San Francisco,1973.

18- Haldiki M, Batistakis Y, Vazirgiannis M. Cluster validity methods, SIGMOD, Record 31, 40-45, 2002.

19- Milligan, G. W. and Cooper MC. An examination of procedures for determining the number of clusters. Psychometrika, 50, 159-179, 1985.

20- Gordon A. D. (1999) Classification, 2nd ed. Chapman and Hall, 1999.

21- Michael J. Crawley. The R Book. Imperial College London at Silwood Park, UK, 2007.

22- The R Project for Statistical Computing Available from: URL: http://www.r-project.org, Version 2.8.