مجله علمی دانشگاه علوم پزشکی سمنان- جلد ۱۲، شماره ۱ (پیاپی ۳۷)، پائیز ۱۳۸۹

کاربرد خوشه‌بندی فازی در تحلیل پروتئین‌های مرتبط با سرطان‌های مری، معده و کلون بر اساس تشابهات تفسیر هستی‌شناسی ژنی

یلدا زرنگارنیا [1*](M.Sc)، حمید علوی‌مجد [2](Ph.D)، مصطفی رضایی‌طاویرانی [2](Ph.D)، نصیبه خیر [3](M.Sc)، علی‌اکبر خادم‌معبودی [2](Ph.D)

۱- *دانشگاه علوم پزشکی شهید بهشتی، شعبه بین الملل*

۲- *دانشگاه علوم پزشکی شهید بهشتی، دانشکده پیراپزشکی، گروه آمار زیستی*

۳- *دانشگاه غیر انتفاعی خاتم*

چکیده

سابقه و هدف: به‌دلیل ایجاد حجم عظیمی از داده‌های پروتئومیکی و نیاز به روش‌های جدید تحلیل نتایج آزمایشگاهی، تحلیل جمعی پروتئین‌ها می‌تواند علاوه بر صرف زمان کم‌تر ما را در شناسایی الگوهای جدید در مجموعه داده‌ها یاری کند. تحلیل خوشه‌ای به عنوان یک روش آماری مطلوب، ابزاری است که می‌تواند در تحلیل این‌گونه داده‌ها مورد استفاده قرار گیرد. هدف از این پژوهش ارزیابی کارایی روش خوشه‌بندی فازی در شناسایی الگوهای جدید در مجموعه پروتئین‌های مرتبط با سرطان‌های دستگاه گوارش بوده است.

مواد و روش‌ها: در این پژوهش پروتئین‌های شناسایی شده مرتبط با سرطان‌های مری، معده و کلون مورد تحلیل خوشه‌بندی فازی قرار گرفته‌اند. بر اساس هر یک از ابعاد هستی‌شناسی ژنی (Gene Ontology) شامل فرآیند بیولوژیکی، جایگاه سلولی و کارکرد مولکولی، به طور جداگانه روش خوشه‌بندی فازی اجرا گردید و نتایج حاصله با هم مقایسه شدند.

یافته‌ها: پس از خوشه‌بندی فازی پروتئین‌ها، مقدار شاخص غیر فازی بر اساس فرآیند بیولوژیکی، جایگاه سلولی و کارکرد ملکولی به ترتیب ۰/۴۱، ۰/۵۵ و ۰/۳۵ به دست آمد که مخصوصاً در مورد خوشه‌بندی بر اساس کارکرد ملکولی نشان‌دهنده مناسبت روش خوشه‌بندی فازی بوده است. با وجود چشم‌گیر نبودن عرض سایه نمای کل خوشه‌بندی‌های حاصل، اکثر پروتئین‌ها در هر خوشه دارای اشتراکات بیولوژیکی قابل توجه شدند. با بکارگیری نرم‌افزار Term Enrichment و تعیین عبارت‌های غنی شده آماری در مجموعه کل داده‌ها و در خوشه‌ها مشخص شد که روش خوشه‌بندی فازی به خوبی توانسته است الگوهای تفسیر جدیدی را در مجموعه داده‌ها آشکار سازد.

نتیجه‌گیری: با بررسی نتایج حاصل از خوشه‌بندی فازی مشخص شد که این روش می‌تواند در جهت تحلیل بهتر و انعطاف‌پذیرتر پروتئین‌ها مورد استفاده قرار گیرد. روش خوشه‌بندی فازی، پروتئین‌هایی را که دارای تشابهات بیش‌تری بوده‌اند با احتمال بیش‌تری در کنار هم قرار داده است، لذا می‌توان از این روش در حالت‌هایی که مشخصه‌های برخی از پروتئین‌ها مجهول می‌باشد، استفاده نمود. هم‌چنین مشخص شد پروتئین‌هایی که بر اساس تشابهات مولفه سلولی در کنار هم قرار می‌گیرند دارای تشابهات بیولوژیکی و عمل‌کردی نیز هستند که این مساله باید مورد بررسی‌های بیش‌تر قرار گیرد.

واژه‌های کلیدی: بیوانفورماتیک، تفسیر هستی‌شناسی ژنی ، خوشه‌بندی فازی، سرطان دستگاه گوارش

مقدمه

امروزه سرطان‌ها به عنوان یکی از مهم‌ترین دلایل مرگ و

* نویسنده مسئول، تلفن: ۰۹۱۲۴۱۸۷۸۰۵ نمابر: ۰۹۱۲۴۱۸۷۸۰۵
Email: y.zarnegar@gmail.com

تاریخ دریافت: ۱۳۸۸/۵/۲۰ تاریخ پذیرش: ۱۳۸۹/۴/۲۹

میر انسان‌ها شناخته می‌شوند. به طوری که طبق آمار سازمان بهداشت جهانی در سال ۲۰۰۰ میلادی، نزدیک به ۱۰ میلیون نفر مورد جدی ابتلا به سرطان و ۶ میلیون نفر مرگ ناشی از سرطان گزارش گردیده است [۱]. سرطان‌های دستگاه گوارش از جمله سرطان‌های شایع در جهان می‌باشند به طوری که به عنوان مثال سرطان معده در حال حاضر به تنهایی نزدیک به ۱۰ درصد کل سرطان‌ها را در جهان تشکیل می‌دهد و یکی از شایع‌ترین انواع سرطان‌ها می‌باشد [۲،۱]. در این میان بدیهی است که شناسایی بیومارکرهای وابسته به بیماری که پیش از ظهور علایم بیماری خود را نشان دهند اهمیت ویژه‌ای خواهد داشت. پروتئین‌ها به دلایل مختلف، بیومارکرهای بسیار خوبی محسوب می‌شوند. با بررسی پروتئین‌ها می‌توان به طور مستقیم عوامل موثر در بیماری را مورد مطالعه قرار داد [۳]. با پیشرفت آزمایش‌های پروتئومیکی با توان بالا مانند آرایه‌بندی پروتئین‌های تصفیه شده این نیاز وجود دارد تا پروتئین‌ها را به صورت جمعی مورد مطالعه قرار دهیم. ابزارهای بسیاری برای تحلیل مجموعه‌های پروتئین‌ها وجود دارند، اما اغلب آن‌ها نتایج حاصله را در یک تصویر ساده و شفاف و قابل تفسیر ارائه نمی‌دهند. PANDORA ابزاری است که مجموعه‌ای از پروتئین‌ها را با توجه به تفاسیر مشترک آن‌ها خوشه‌بندی و نتایج را به صورت یک نمودار نمایش می‌دهد. SGD (Saccharomyces Genome Database) به کمک ابزاری مانند GO Term FINDER و GO Slim Mapper و GO Annotation Summary برای اجتماع مخمر، هر پروتئین و تمام اثرات متقابل آن با دیگر پروتئین‌ها را مورد تحلیل قرار می‌دهد. Web Gestalt ابزاری است که به کمک آن می‌توان مجموعه مورد علاقه از پروتئین‌ها را وارد کرد و بیش از بیست نوع تفسیر را برای بکارگیری تشخیص داد [۴]. هونگ‌بین شن و همکاران در سال ۲۰۰۵ از روش جدید خوشه‌بندی فازی با سرپرست، برای پیش‌بینی کلاس‌های ساختاری پروتئین‌ها استفاده نمودند. بر اساس این روش در مرحله آموزش اطلاعات مربوط به کلاس‌های ساختاری پروتئین‌ها مورد استفاده قرار گرفت [۵]. پوپسکو و همکاران در سال ۲۰۰۴ به منظور یافتن عبارت‌هایی که بتوانند در خوشه‌بندی بر اساس تشابهات هستی‌شناسی ژنی نمایندگان خوبی برای خوشه‌ها باشند از روش خوشه‌بندی سلسله مراتبی و اندازه مشابهت‌های حاصل از BLAST و اندازه مشابهت فازی استفاده نمودند [۶]. هوگو و همکاران در سال ۲۰۰۶ به منظور بررسی کارایی خوشه‌بندی پروتئین‌ها در تسریع مطالعه آن‌ها، پروتئین‌ها را به کمک اندازه مشابهت‌های حاصل از BLAST خوشه‌بندی نمودند. پس از بررسی خوشه‌ها از لحاظ هستی‌شناسی ژنی، نتیجه‌گیری شد که مرکز هر خوشه می‌تواند اطلاعات پروتئین‌های خوشه را در برگیرد [۷]. کریستین اواسکا و همکاران در سال ۲۰۰۸ با اندازه‌گیری مشابهت‌های بین عبارات هستی‌شناسی ژنی و اجرای روش خوشه‌بندی سلسله مراتبی، توانستند روشی ارائه دهند که در شناسایی سریع ژن‌هایی که دارای عبارات GO مشترک هستند کمک نماید [۸]. در این پژوهش روش خوشه‌بندی فازی به منظور تحلیل پروتئین‌های مشترک مرتبط با سرطان‌های دستگاه گوارش، مورد استفاده قرار گرفته و هدف این بوده است که آیا می‌توان به کمک خوشه‌بندی پروتئین‌ها بر اساس تشابهات تفاسیر هستی‌شناسی ژنی آن‌ها، به الگوهایی در زیر مجموعه‌هایی از پروتئین‌ها دست یافت که در مطالعه جداگانه آن‌ها، قابل تشخیص نباشند.

مواد و روش‌ها

در این پژوهش از داده‌های جمع‌آوری شده در مرکز تحقیقات پروتئومیک دانشکده پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی استفاده شده است. در این داده‌ها، تعداد ۱۷ پروتئین به عنوان پروتئین‌های دخیل مشترک در سرطان‌های مری، معده و کلون شناسایی شده‌اند [۹]. اطلاعات مورد نیاز برای تحلیل این ۱۷ پروتئین بر اساس هستی‌شناسی ژنی از وب‌سایت هستی‌شناسی ژنی جمع‌آوری شد [۱۰].

پروژه GO سه دسته واژگان کنترل شده 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 مورد بررسی قرار داد. این سه ویژگی عبارتند از فرآیند بیولوژیکی (Biological Process, BP)، جایگاه سلولی (Cellular-Component, CC) و عملکرد ملکولی (Molecular Function, MF). لذا براساس هر یک از این وجوه، گرافهای GO از وبسایت Gene Ontology استخراج شدند. برای اندازهگیری تشابهات بین پروتئینها جهت بکارگیری در خوشهبندی آنها، از میان روشهای موجود از روش simUI استفاده گردید [۱۱]. بر اساس 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. به عنوان مثال تشابه گرافیکی بین گراف مربوط به تفاسیر بیولوژیکی پروتئین پنجم (Osteonectin) و پروتئین دوازدهم (Annexin-A2) برابر است با تعداد گرههای مشترک به تعداد ۷ گره، تقسیم بر کل ۱۸ گره=۰/۳۹. لذا عدم تشابه برابر خواهد بود با: ۱-۰/۳۹ [۱۲].

روشهای خوشهبندی شامل روشهای خوشهبندی سلسله مراتبی و غیر سلسله مراتبی هستند. در روشهای غیر سلسله مراتبی (Partitioning methods) تعداد k خوشه ساخته میشود و دادهها به k گروه تقسیم میشوند به طوری که هر گروه شامل حداقل یک عنصر باشد. یکی از روشهای غیر سلسله مراتبی روش خوشهبندی فازی است که در این پژوهش مورد استفاده قرار گرفته است. این روش بر اساس اصل Fuzziness میباشد. در واقع در یک افراز قطعی هر عنصر فقط و فقط به یک خوشه تعلق خواهد گرفت. اما در خوشهبندی فازی ما قطعیت نداریم و به دنبال تعیین ضرایب عضویت یا احتمال عضویت هر عنصر در هر خوشه هستیم. روش خوشهبندی فازی در شرایط مبهم که به دلیل وجود عناصری که از لحاظ تعلق به خوشهها در وضعیت بینابین هستند و خوشهها دچار همپوشانی میشوند، کارایی بیشتری دارد [۱۴،۱۳]. به کمک شاخص غیر فازی میتوان کارایی خوشهبندی فازی را مورد ارزیابی قرار داد. همچنین میتوان پس از اجرای روش خوشهبندی با تعیین نزدیکترین خوشهبندی قطعی، یک نمایش گرافیکی از خوشهها داشت که آنرا Silhouette یا سایهنما مینامیم. میتوان به ازای هر خوشهبندی یک نمودار سایهنما داشت و با کنار هم قرار دادن آنها، کیفیت خوشهبندیها را با هم مقایسه نمود. محور افقی این نمودار نشاندهنده عرض خوشهها و محور عمودی نحوه تعلق گرفتن عناصر به خوشهها را نشان میدهد.

یکی از مسایل 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 وجود نداشته باشد. مقادیر بزرگ K نیز خوشههای کوچکی را میدهد که ممکن است اطلاع مناسبی را در اختیار قرار ندهند، زیرا روابط بین تعداد کمتری از اقلام را نشان میدهند. لذا به منظور محاسبه تعداد بهینه خوشهها، از روش Silcheck در Bioconductor بر اساس بیشینه نمودن متوسط عرض سایه نمای کل و با تعیین تعداد حداکثر ۵ خوشه، استفاده گردید [۱۳،۱۲]. به منظور ارزیابی و اعتبارسنجی روش خوشهبندی فازی بر اساس تشابهات تفاسیر هستیشناسی ژنی، نتایج خوشهبندی را با یک روش تحلیلی مورد استفاده در تحلیل ریز آرایه DNA، با کمک تعیین عبارتهای GO بیش نشان داده شده آماری مقایسه میکنیم. بدین صورت که تفسیر GO یک زیر مجموعه از ملکولها با تفسیر GO مجموعه مرجع (UniProtKB) مقایسه میشود [۱۵،۱۲]. اگر هر عبارت هستیشناسی ژنی یا عبارتهای نیایی آن (عبارت نیایی، والد عبارت مورد نظر ما در گراف GO میباشد.) بر اساس BP، CC و یا MF بیشتر از حد معمول در هر زیر مجموعه نسبت به مجموعه مرجع رخ دهد، گوییم آن عبارت GO غنی شده آماری است. اگر تفسیر GO پروتئینی که در هر خوشه بیشترین احتمال تعلق به خوشه را دارد یک عبارت غنی شده آماری برای پروتئینهای آن خوشه باشد آنگاه میفهمیم که تفسیر GO پروتئینی که بیشتری احتمال تعلق به خوشه را داشته است، به درستی تفسیرGO پروتئینها را در آن خوشه نشان میدهد. به منظور سنجش توانایی روش خوشهبندی فازی، در آشکارسازی الگوهای

تفسیر جدید در مجموعه داده‌ها، عبارت‌های GOپروتئین‌هایی که در هر خوشه بیش‌ترین احتمال تعلق به خوشه را داشته‌اند را با عبارت‌های غنی شده آماری در کل مجموعه داده‌ها مقایسه می‌کنیم. چنان‌چه عبارات GO پروتئین‌هایی که در هر خوشه بیش‌ترین احتمال تعلق به خوشه را داشته‌اند، عبارات غنی شده آماری در کل مجموعه داده‌ها باشند، آنگاه در واقع به اطلاعات جدیدی در مجموعه داده‌ها دست نیافته‌ایم. اما اگر این عبارات در مجموعه کل داده‌ها غنی شده آماری نباشند، آن‌گاه روش خوشه‌بندی فازی توانسته است که الگوهای تفسیر جدیدی را در مجوعه داده‌ها آشکار نماید. برای اجرای الگوریتم خوشه‌بندی از نرم‌افزار R استفاده گردید [۱۶] و Package های مورد نیاز نیز از وب‌سایت Bioconductor دانلود شدند [۱۷].

نتایج

امتیازهای مشابهت به کمک روش simUI برای هر ۱۷ پروتئین و بر اساس هر یک از وجوه GO به طور جداگانه محاسبه گردید. تعداد بهینه خوشه‌ها برای داشتن خوشه‌بندی قطعی نزدیک به احتمالات حاصل از اجرای روش فازی، بر اساس بیشینه نمودن عرض سایه‌نما به ترتیب تعداد ۵ خوشه بر اساس فرایند بیولوژیکی (BP)، عملکرد ملکولی (MF) و مولفه سلولی (CC) تعیین شد. نتایج حاصل از خوشه‌بندی به روش فازی را می‌توان در جدول (۱) مشاهده نمود. در این جدول احتمال تعلق هر یک از پروتئین‌ها به خوشه‌ها بر اساس هر یک از وجوه هستی‌شناسی ژنی قابل مشاهده است. نمودارهای سایه‌نمای مربوط به نزدیک‌ترین خوشه‌بندی قطعی یا سخت بر اساس احتمالات حاصل شده را می‌توان در شکل‌های (۱) و (۲) و (۳) مشاهده نمود. عبارت GO پروتئین‌هایی را که در هر خوشه دارای بالاترین احتمال تعلق به آن خوشه بوده‌اند در برابر شماره آن‌ها و عرض سایه نمای کل در پایین هر خوشه بندی نوشته شده است. عرض سایه‌نمای کل در پایین هر خوشه‌بندی نوشته شده است. عرض سایه‌نما نیز در مقابل هر خوشه به همراه تعداد پروتئین‌ها تعلق گرفته به آن خوشه قرار گرفته است. شاخص غیر فازی برای خوشه‌بندی بر اساس فرآیند بیولوژیکی، جایگاه سلولی و عملکرد ملکولی به ترتیب ۰/۴۱، ۰/۵۵ و ۰/۳۵ به دست آمد. می‌توان گفت که خوشه‌بندی فازی خصوصاً بر اساس مولفه سلولی عملکرد خوبی داشته است. برای خوشه‌بندی فازی، ساختارهای خوشه‌بندی قطعی با توجه به احتمالات به دست آمده دارای مقادیر S_i^D چشم‌گیری نبوده‌اند. این ساختارها بر اساس S_i^D (BP =0.23) S_i^D =0.20,MF ضعیف بوده‌اند و مقدار عرض سایه‌نمای کل در خوشه‌بندی بر اساس CC (S_i^D CC= 0.34) نیز مقدار کوچکی شده است.

مقدار S_i^D برای خوشه‌بندی بر اساس BP مقداری ضعیف و در کل تعداد ۳ مقدار از ۵ مقدار S_i^D مقادیر ضعیفی شده‌اند. در خوشه اول با وجود عرض خوشه کوچک (S_i^D =0.11) مشخص شد که پروتئین‌های شماره ۴ و۳ در اثرات متقابل بین ارگانیسم‌ها و پروتئین‌های شماره ۴و ۸ نیز در حرکت سلول نقش دارند. خوشه دوم نیز مقدار عرض خوشه کوچکی (S_i^D =0.13) دارد، اما چهار پروتئین شماره ۲ و ۱۰و ۱۵ و۱۶ موجود در این خوشه در تنظیم apoptosis نقش دارند. خوشه سوم (S_i^D =0.35 BP)خوشه نسبتاً متعادلی است. بررسی بیش‌تر نشان می‌دهد که پروتئین‌های شماره ۵ و ۱۲ در توسعه سیستم اسکلتی و پروتئین ۱۷ نیز در توسعه اندام ماهیچه‌ای نقش دارند. به علاوه هر سه این پروتئین‌ها متصل شوندگان به یون کلسیم هستند. خوشه چهارم نیز دارای عرض خوشه (S_i^D =0.32) کوچکی است و هر دو پروتئین درون این خوشه تنظیم‌کننده‌های مرگ سلولی (Apoptosis) هستند. هر دو پروتئین خوشه پنجم نیز با وجود عرض خوشه کوچک، در فرایند بیولوژیکی پاسخ التهابی نقش دارند. لذا روش فازی با وجود نتیجه دادن عرض خوشه کل کوچک بر اساس BP به خوبی توانسته پروتئین‌ها را از یک‌دیگر تمیز دهد.

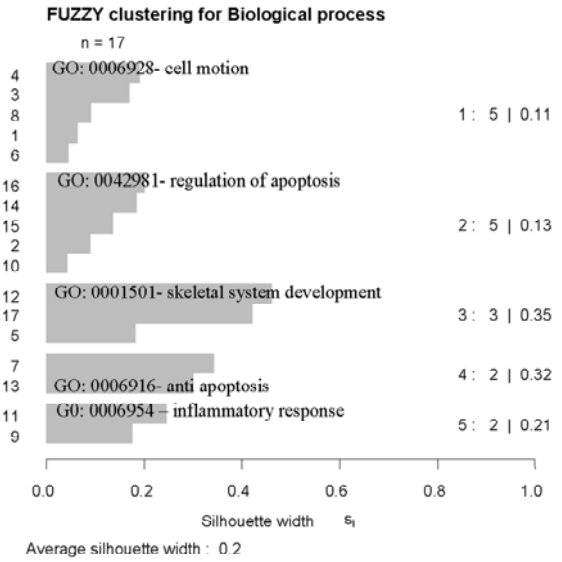
جدول ۱. ضرایب عضویت پروتئین‌ها در خوشه‌ها بر اساس احتمالات به دست آمده از اجرای روش خوشه بندی فازی

عملکرد مولکولی					مولفه سلولی					فرآیند بیولوژیکی					اسامی استاندارد پروتئین‌ها	
خوشه‌ها					خوشه‌ها					خوشه‌ها						
۵	۴	۳	۲	۱	۵	۴	۳	۲	۱	۵	۴	۳	۲	۱		
ضرایب عضویت در خوشه‌ها (به درصد)					ضرایب عضویت در خوشه‌ها (به درصد)					ضرایب عضویت در خوشه‌ها (به درصد)						
۱۰	۱۸	۲۷	۷	۳۸	۲۱	۱۴	۱۸	۱۸	۲۹	۲۱	۱۶	۱۱	۱۷	۳۶	CAH2	۱
۸	۱۷	۲۳	۷	۴۵	۸	۱۱	۲۳	۱۵	۴۴	۹	۱۷	۵	۵۹	۹	SODM	۲
۲	۱۰	۵	۷۹	۴	۱	۱	۲	۹۳	۳	۸	۶	۵	۷	۷۴	K2C8	۳
۰	۱	۰	۹۹	۰	۱	۱	۲	۹۳	۳	۵	۴	۳	۴	۸۴	VIME	۴
۷	۱۰	۱۱	۴	۶۷	۱	۲	۹۱	۲	۴	۱۸	۲۰	۲۷	۱۷	۱۸	SPRC	۵
۰	۱	۰	۹۹	۰	۰	۱۰۰	۰	۰	۰	۲۱	۱۹	۱۲	۱۸	۳۱	DESM	۶
۹	۲۳	۳۸	۱۰	۱۹	۱۰۰	۰	۰	۰	۰	۴	۸۳	۲	۶	۳	PRDX2	۷
۹	۲۶	۳۰	۱۷	۱۸	۳	۴	۸	۸	۷۸	۱۶	۱۳	۱۰	۱۳	۴۷	ACTB	۸
۱۰	۲۴	۳۹	۱۰	۱۷	۲	۳	۸۱	۴	۱۰	۷۸	۸	۴	۵	۵	A1AT	۹
۵	۵۱	۱۵	۱۷	۱۲	۶	۷	۱۳	۴۴	۲۹	۷	۱۷	۴	۶۴	۷	HSPB1	۱۰
۱۰	۱۵	۱۵	۶	۵۳	۲	۳	۸	۷	۸۱	۸۴	۵	۳	۴	۵	S10A9	۱۱
۱۰۰	۰	۰	۰	۰	۷	۱۰	۶۰	۱۰	۱۳	۱	۱	۹۷	۱	۱	ANXA2	۱۲
۱۰۰	۰	۰	۰	۰	۱۰۰	۰	۰	۰	۰	۶	۸۵	۲	۶	۳	ANXA5	۱۳
۹	۲۱	۴۳	۹	۱۷	۳	۴	۱۰	۸	۷۴	۱۱	۱۲	۶	۶۰	۱۰	PCNA	۱۴
۱۱	۲۱	۳۲	۹	۲۶	۷	۱۰	۳۵	۱۳	۳۵	۱۵	۱۸	۹	۴۳	۱۵	CALR	۱۵
۶	۴۶	۲۰	۱۳	۱۵	۸	۱۲	۲۳	۱۴	۴۳	۵	۸	۳	۷۹	۵	PHB	۱۶
۶	۵۱	۱۷	۱۴	۱۳	۱۰۰	۰	۰	۰	۰	۱	۱	۹۶	۱	۱	TAGL	۱۷
۱۷					۱۷					۱۷					تعداد پروتئین‌های خوشه بندی شده	

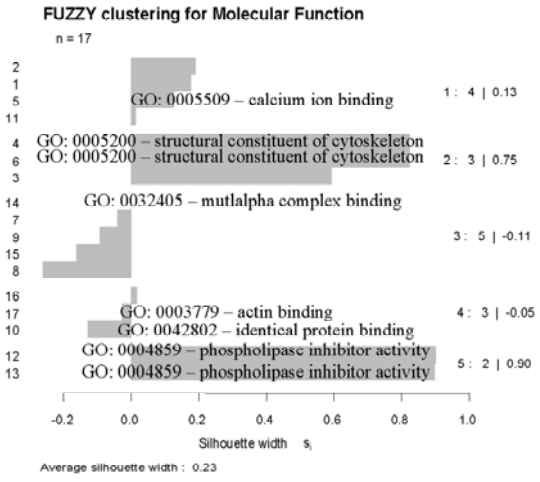
مقدار S_i^D برای خوشه‌بندی بر اساس MF نیز مقداری کم‌تر از ۰/۲۵ شده است که نشان می‌دهد ساختار خوشه‌بندی چشم‌گیری به دست نیامده است. با وجود عرض کوچک خوشه اول (S_i^D =0.13) پروتئین‌های شماره ۵ و ۱۱ موجود در این خوشه متصل شوندگان به یون‌های کلسیم هستند و پروتئین شماره ۱ نیز متصل‌شونده به یون روی می‌باشد و لذا به درستی در کنار هم قرار گرفته‌اند. خوشه دوم عرض سایه‌نمای خوبی (S_i^D =0.75) داشته است و بررسی‌ها نشان می‌دهد پروتئین‌های شماره ۴ و۳ و ۶ بسیار به هم شبیه و هم خانواده هستند و به خوبی از بقیه پروتئین‌ها جدا شده‌اند. هم‌چنین پروتئین‌های هم خانواده شماره ۱۲ و۱۳ که متصل‌شوندگان به یون کلسیم هستند، نیز در خوشه پنجم با عرض سایه‌نمای (S_i^D =0.90) و با احتمال ۱۰۰٪ در کنار هم قرار گرفته‌اند. در خوشه سوم و چهارم نیز عرض‌های سایه‌نما منفی شده که با توجه به کوچکی احتمالات تعلق پروتئین‌های این خوشه‌ها، می‌توان گفت که آن‌ها از بقیه پروتئین‌ها متفاوت بوده‌اند.

خوشه بندی بر اساس CC با $S_i^D = 0.34$ ساختار خوشه بندی نسبتا خوبی بوده است. در خوشه اول پروتئین‌های شماره ۸ و ۱۱ و۱۴ دارای بیش‌ترین احتمال تعلق به خوشه بوده‌اند که هر سه این پروتئین‌ها در نوکلئوز فعالیت می‌کنند و لذا به درستی در کنار هم قرار گرفته‌اند. هم‌چنین پروتئین شماره ۸ متصل‌شونده به آنزیم و به ATP، پروتئین شماره ۱۱ متصل‌شونده به یون کلسیم و پروتئین

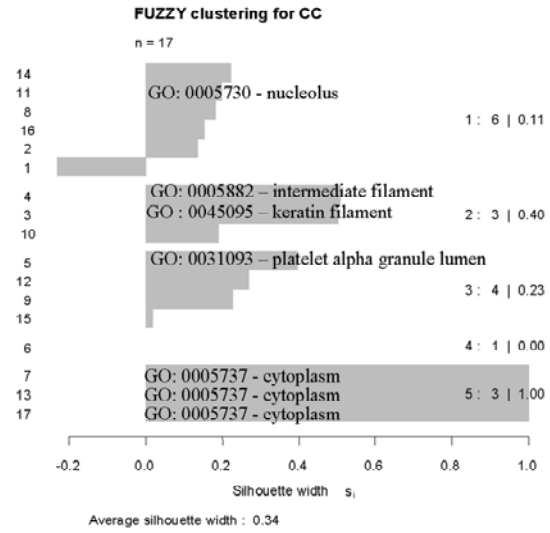
شماره ۱۴ نیز متصل‌شونده به DNA می‌باشد. پروتئین‌های ۷و ۱۳ و۱۷ در خوشه پنجم با احتمال ۱۰۰٪ در کنار هم‌دیگر قرار گرفته‌اند. هر سه این پروتئین‌ها در سیتوپلاسم فعالیت دارند و پروتئین‌های شماره ۱۳و ۱۷ متصل‌شوندگان به یون کلسیم هستند. پروتئین‌های شماره۱۰ و ۱۵ با احتمالی کم به خوشه‌های دوم و سوم قرار گرفته‌اند. اما پروتئین‌های شماره ۳ و۴ و پروتئین‌های شماره ۵ و ۹ و ۱۲ با احتمال‌های بالاتری در خوشه‌های دوم و سوم در کنار هم قرار گرفته‌اند.



شکل۱. نمودار سایه‌نما برای خوشه بندی سخت متناظر با خوشه بندی فازی با توجه به ضرایب عضویت حاصل از اجرای روش خوشه بندی فازی بر اساس فرآیند بیولوژیکی



شکل۲. نمودار سایه‌نما برای خوشه بندی سخت متناظر با خوشه بندی فازی با توجه به ضرایب عضویت حاصل از اجرای روش خوشه بندی فازی بر اساس کارکرد مولکولی



شکل۳. نمودار سایه‌نما برای خوشه بندی سخت متناظر با خوشه بندی فازی با توجه به ضرایب عضویت حاصل از اجرای روش خوشه بندی فازی بر اساس جایگاه سلولی

برای ارزیابی اعتبار خوشه‌بندی‌ها به روش فازی بر اساس مشابهت تفاسیر GO، عبارات غنی شده آماری خوشه‌ها به کمک نرم‌افزار Term Enrichment تعیین شدند. از کل تعداد ۱۵ خوشه ایجاد شده به روش فازی تعداد ۷ خوشه دارای یک یا بیش‌تر عبارات غنی شده آماری بودند. مقایسه این عبارات با عبارت‌های GO پروتئین‌هایی که بیش‌ترین احتمال را در تعلق به خوشه‌ها به خود اختصاص داده بودند نشان داد که عبارت‌های GO مربوط به ۱۰۰٪ پروتئین‌هایی که در هر خوشه بالاترین احتمال تعلق به خوشه‌ها را داشته‌اند (۷ خوشه از ۷ خوشه)، غنی شده برای آن خوشه‌ها هستند. این نشان‌دهنده درستی کارکرد روش خوشه‌بندی فازی در خوشه‌بندی پروتئین‌ها بر اساس مشابهت‌های تفاسیر هستی‌شناسی ژنی آن‌هاست.

به منظور بررسی توانایی روش خوشه‌بندی فازی در شناسایی الگوهای جدید، در بررسی مجموعه داده‌ها پیش از خوشه‌بندی، Term Enrichment توانست عبارات تنها ۰/۲۹ (۶ پروتئین از ۲۱ پروتئین) از پروتئین‌های دارای بیش‌ترین احتمال تعلق به خوشه‌ها را به عنوان غنی‌شده آماری شناسایی کند. لذا روش فازی در خوشه‌بندی پروتئین‌ها بر اساس مشابهت‌های تفاسیر GO آن‌ها، توانسته است تفسیر بهتری از

پروتئین‌ها را در اختیار ما بگذارد و به عنوان ابزاری مفید در ۳ خوشه (خوشه‌های سوم و چهارم مربوط به BP و خوشه اول مربوط به MF) الگوهای تفسیر جدیدی را که قبلاً مشخص نشده‌اند، شناسایی کند.

در کل می‌توان گفت روش خوشه‌بندی فازی بر اساس تفاسیر GO، علی‌رغم ایجاد عرض سایه‌نمای کل ضعیف، به خوبی توانسته است الگوهای تفسیر جدیدی را شناسایی کند که متفاوت از عبارت‌های بیش نشان داده شده آماری در مجموعه کل داده‌ها بوده‌اند. می‌توان گفت اگر چه بر اساس منابع موجود در استفاده از روش فازی خوشه‌های به وجود آمده با $S_i^C \leq 0.25$ چندان قابل تفسیر نیستند اما چنین چیزی در مورد خوشه‌بندی پروتئین‌ها درست نیست و خوشه‌های با مقادیر کم S_i^C نیز خوشه‌های قابل تفسیر و ارزشمند از لحاظ تفسیرهای بیولوژیکی هستند.

بحث و نتیجه‌گیری

در مطالعه مجموعه‌ای از پروتئین‌ها هدف آن است تا به الگوها و روابط و تفاسیری در بین آن‌ها دست یابیم که شاید ناشناخته مانده باشند. کاپلان و لینیال در سال ۲۰۰۵ فاصله بین هر دو پروتئین را به عنوان تابعی از تعداد عبارت‌های مشترک در تفسیر هستی‌شناسی ژنی آن دو تعیین نمودند، به طوری‌که عبارت‌های کم‌تر رایج مانند پروتین های شوک حرارتی (Heat shock protein) امتیاز بالاتری نسبت عبارات بیش‌تر رایج مثل آنزیم می‌گرفتند. آن‌ها خوشه‌بندی سلسله مراتبی موفقی را انجام دادند که در آن خوشه‌ها هیچ‌گونه تفسیر اشتباهی نداشتند [۱۸]. شریل ولتینگ و همکاران در سال ۲۰۰۶ مجموعه‌ای از داده‌های پروتئینی مربوط به مخمر را به کمک روش خوشه‌بندی تقسیم‌بندی حول نماینده‌ها (PAM) بر اساس تشابهات هستی‌شناسی ژنی حاصل از روش simUI خوشه‌بندی نموده و در داخل خوشه‌ها به الگوهای تفسیر جدید دست یافتند [۱۲]. آن‌ها علاقه‌مند بودند که از روش PAM در خوشه‌بندی پروتئین‌های موجودات عالی‌تر استفاده کنند. در پژوهش‌های یاد شده، هر پروتئین به صورت قطعی به خوشه‌ای تعلق می‌گرفت و موقعیت‌های خاص، هم‌چون وجود پروتئین‌هایی که از لحاظ تعلق به خوشه‌ها دارای وضعیت بینابین باشند، پیش‌بینی نشده بود. در این پژوهش با اجرای روش خوشه‌بندی فازی برای تحلیل پروتئین‌های مرتبط با سرطان‌های دستگاه گوارش نشان داده شد که تحلیل فازی علاوه بر کمک در شرایط مبهم هم‌پوشانی خوشه‌ها، توانسته است الگوهای تفسیر جدیدی را در مجموعه داده‌ها آشکار سازد. با اجرای روش خوشه‌بندی فازی و تعیین عبارت‌های غنی شده آماری مشخص شد که روش خوشه‌بندی فازی توانسته است زیر مجموعه‌های کوچکی را در داخل داده‌ها ایجاد کند و با تعیین عبارت GO پروتئین‌هایی که در هر خوشه بیش‌ترین احتمال تعلق را داشته‌اند، به الگوهای جدیدی دست یابد که شاید با نگاه کلی به داده‌ها به چشم نیامده‌اند. این الگوهای شناسایی شده ما را به مطالعات و پیگیری آزمایشات بیش‌تر جهت بررسی مکانیسم‌ها و تاثیر وجود اثرات متقابل در غربال‌گری آرایه پروتئین‌ها ترغیب می‌کند. در روش خوشه‌بندی فازی ۷ عبارت GO غنی شده آماری برای پروتئین‌هایی که در ۷ خوشه (که عبارات غنی‌شده آماری داشتند)، دارای بالاترین احتمال تعلق بودند شناسایی شدند که در بررسی کل داده‌ها توانستیم ۴ عبارت از ۷ عبارت را به عنوان غنی‌شده آماری شناسایی کنیم که این خود توانایی روش خوشه‌بندی فازی را در شناسایی الگوهای جدید در مجموعه داده‌ها نشان می‌دهد. بنابراین لازم است که خوشه‌های سوم و چهارم مربوط به BP و خوشه اول مربوط به MF که در نتیجه اجرای روش خوشه‌بندی فازی حاصل شده‌اند، مورد بررسی و مطالعه بیش‌تر قرار گیرند. هم‌چنین بنابر نتایج به دست آمده از خوشه‌بندی فازی، پروتئین‌های شماره ۳ و۴ در هر سه خوشه‌بندی در کنار یک‌دیگر قرار گرفته‌اند و باید مورد بررسی بیش‌تر قرار گیرند. می‌توان گفت خوشه‌بندی فازی به خوبی توانسته است پروتئین‌های مشابه را با دادن احتمالات بالا در کنار هم قرار دهد و به پروتئین‌هایی که کم‌تر به دیگر پروتئین‌های خوشه شبیه بوده‌اند احتمال کم‌تری اختصاص

دهد. بررسی بیش‌تر خوشه‌های حاصل از خوشه‌بندی فازی بر اساس CC نشان داد که پروتئین‌های قرار گرفته در هر خوشه از لحاظ عمل‌کرد و فرآیندهای بیولوژیکی که در آن‌ها نقش دارند دارای تشابهات بسیاری هستند که خود قابل تامل است. شاید با مطالعات بیش‌تر بتوان گفت که چنان‌چه مجموعه‌ای از پروتئین‌ها را داشته باشیم که مکان آن‌ها در سلول مشخص باشد با خوشه‌بندی آن‌ها بر اساس CC، بتوان خوشه‌هایی را تولید کرد که پیش‌بینی کنند پروتئین‌های موجود در هر خوشه دارای کارکردهای مشترکی هستند و یا در فرآیندهای بیولوژیکی مشابهی نقش دارند.

منابع

[1] 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.

[2] Parkin DM. Epidemiology of cancer: global patterns and Trends. Toxicol Lett 1998; 102-10: 227-234.

[3] Khatri P. and Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems, Bioinfomatics 2005; 21: 3587-3595.

[4] Rezaei Tavirani M, Marashi A, Ghalanbar F. and Mostafavi M. Preoteomics. Andishe Zohoor Publication, 1384. (Persian).

[5] Shen HB, Yang J, Liu XJ. and Chou KC. Using supervised fuzzy clustering to predict protein structural classes. Biochem Biophys Res Commun 2005; 334: 577-581.

[6] 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.

[7] Hugo Bastos, Daniel Faria, Catia pesquita and Andreo Flacao. Using GO terms to evaluate protein clustering. In ISMB/ECCB 2007 SIG Meeting Program Materials, 2007; Pages 107—110.

[8] Ovaska K, Laakso M. and Hautaniemi S. Fast Gene Ontology based clustering for microarray experiments. BioData Min 2008; 1: 11.

[9] Khaier N, Rezaei Tavirani M. and Rostami A. 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. (Persian).

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

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

[12] Wolting C, McGlade CJ. and Tritchler D. Cluster analysis of protein array results via similarity of Gene Ontology Annotation. BMC Bioinformatics 2006; 7: 338.

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

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

[15] Datta S. and Datta S. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. BMC Bioinformatics 2006; 7: 397.

[16] Michael J Crawley; The R Book, Imperial College London at Silwood Park, UK,2007.

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

[18] Kaplan N, Vaaknin A. and Linial M. PANDORA: keyword-based analysis of protein sets by integration of annotation sources. Nucleic Acids Res 2003; 31: 5617-5626.

Application of fuzzy clustering in analysis of included proteins in esophagus, stomach and colon cancers based on similarity of Gene Ontology annotation

Yalda Zarnegarnia (M.Sc)[*1], Hamid Alavi Majd (Ph.D)[2], Mostafa Rezaei Tavirani (Ph.D)[2], Nasibe Khaier (M.Sc)[3], Ali akbar Khadem Maboodi (Ph.D)[2]

1 - The International Branch of Shahid Beheshti Medical Science University, Tehran, Iran

2 – Dept. of Biostatistics, Paramedical Sciences Faculty, Shahid Beheshti Medical Science University, Tehran, Iran

3 - Khatam Non-profit university , Tehran, Iran

(Received: 11 Agu 2009 Accepted: 20 Jul 2010)

Introduction: Because of producing large amount of proteomics data and requiring new procedures for analyzing them, collective analysis of proteins can help us in identifying new annotation patterns in dataset. Furthermore, this type of analysis is a time- consuming process too. Cluster analysis, as a suitable statistic procedure, can be used for analyzing these datasets. This paper's objective was evaluating the efficiency of fuzzy clustering method in recognizing new patterns within proteins which are related to gastric cancers.

Materials and Methods: Fuzzy clustering procedure has been used to analyze the identified included proteins in esophagus, stomach and colon cancers. Proteins were clustered based on three aspects of Gene Ontology (GO) and results were compared.

Results: Fuzzy clustering was implemented and non-fuzziness indexes based on biological process, cellular component and molecular function were obtained equal to 0.41, 0.55 and 0.35, respectively. Obtained index based on molecular function showed the efficiency of fuzzy clustering method. Despite of non-substantial silhouette widths for the entire dataset, most of the proteins in each cluster had remarkable biological communions. Using Term Enrichment software to determine statistically enriched GO terms in the entire dataset and clusters, it was cleared that the fuzzy clustering has revealed novel annotation patterns within dataset that would not have been identified otherwise.

Conclusion: Considering fuzzy clustering outputs, the efficiency of this method for better and flexible proteins analysis was cleared. As fuzzy clustering method has placed proteins, that have more similarities, with high probabilities together. Therefore, it can be used for the situations that some of proteins have unknown characteristics. Furthermore it seems that the proteins clustered via their cellular component similarities, have also biological and functional similarities which this requires more investigations.

Key words: Bioinformatics, Gene Ontology annotation, Fuzzy clustering, Gastric system cancer

***** Corresponding author: Fax: +98 9124187805; Tel: +98 9124187805
y.zarnegar@gmail.com