DOR :20.1001.1.2538161.1399.5.4.2.9

تاریخ دریافت مقاله: ۹۸/۱۰/۲۳
تاریخ پذیرش مقاله: ۹۹/۰۱/۱۴

بهبود تشخیص کمپلکس‌های پروتئینی مبتنی بر خوشه بندی دوگانه داده‌های بیان ژن

امیر لکی زاده*
گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه قم
پست الکترونیکی: lakizadeh@qom.ac.ir

چکیده

با توجه بـــه نقش کمپلکس‌هـــای پروتئینـــی در انجام بســـیاری از کارکردهای ســـلولی موجودات زنده، کشف آن‌هـــا می‌تواند بـــه درک عمیق‌تر ســـازوکارهای تنظیمی ســـلول، توصیف فرآیند تکامل در ســـیگنال‌های ســـلولی، پیش‌بینی عملکرد زیستی پروتئین‌های کشف‌شده و از همه مهم‌تر تحقق اهداف درمانی(تشـــخیص بیماری و طراحی دارو) منجر شود. رویکردهای محاسباتی ارائه‌شده تاکنون به منظور تشـــخیص کمپلکس‌های پروتئینی، به طور عمده بر خوشـــه‌بندی شبکه برهم‌کنش پروتئین–پروتئین تمرکز دارند و این در حالی است که شبکه‌های برهم‌کنش، علاوه بر ایـــن که نوفه‌دار می‌باشـــند، به‌تنهایی، فاقد ســـازوکار لازم بـــرای در نظر گرفتن ماهیت پویای ســـلول در فرآیند تشخیص کمپلکس‌های پروتئینی می‌باشند.

در ایـــن مقالـــه، مدلی ســـه لایـــه مبتنی بـــر رویکرد خوشه‌بندی دوگانه برای تشخیص کمپلکس‌های پروتئینی از منابع داده‌ای مختلف ارائه شده است. لایه‌های اول، دوم و سوم مدل پیشنهادی، به ترتیب وظیفه پویاسازی، کاهش نوفه و تشـــخیص نهایی کمپلکس‌های پروتئینی را بر عهده دارنـــد. مجموعـــه ارزیابی‌های مختلف، نشـــان می‌دهد که روش پیشنهادی توانسته است با استفاده از منابع داده‌ای مختلف در لایه‌های سه‌گانه، ضمن مدیریت بهتر چالش‌های اصلی مسئله مانند نوفه‌دار بودن منابع داده‌ای و ضرورت مدل‌سازی پویای فرآیند تشخیص به کمک داده های بیان ژن، دقت فرآیند تشـــخیص کمپلکس‌هـــای پروتئینی را بر اساس سنجنده‌های مرتبط و در شرایط گوناگون، به میزان قابل توجهی بهبود دهد.

واژه‌های کلیدی: کمپلکس پروتئینی، شـــبکه برهم‌کنش پروتئین–پروتئین، داده‌های بیان ژن، خوشه‌بندی دوگانه، شباهت معنایی.

۱. مقدمه

اکثر پروتئین‌ها عملکردهای زیســـتی خود را در سلول زنـــده در قالب تشـــکیل کمپلکس‌ها انجـــام می‌دهند[۱]. در حقیقت بسیاری از پروتئین‌ها به‌عنوان اجزای کمپلکس‌های ضـــروری و پایدار وجود دارند. برای مثال مولکول حیاتی هموگلوبین، یک ترکیب پایـــدار از چهار پروتئین گلبول‌دار مشخص می‌باشد. هم‌چنین بســـیاری از آنزیم‌ها که نقش راه‌انـــداز را در بســـیاری از واکنش‌های درون‌ســـلولی به عهده‌دارند، مجموعه‌ای از پروتئین‌ها می‌باشـــند که به‌طور هم‌زمان به یکدیگر متصل شـــده و ســـاختار موردنظر را تشـــکیل می‌دهند. علاوه بر آن برهم‌کنش‌هـــای ناپایدار و

* نویسندهٔ مسئول

موقت مانند اثرات هورمون و یا اثرات یک سیگنال موقت، مستلزم تشکیل کمپلکس‌های پروتئینی می‌باشد[۲]. از لحاظ زیستی، کمپلکس‌های پروتئینی اجزاء مولکولی کلیدی برای انجام بسیاری از عملکردهای زیستی از قبیل رونوشت DNA، ترجمه mRNA، انتقال سیگنال سلولی، چرخه سلولی و غیره می‌باشد. برای مثال کمپلکس RNA پلیمراز در فرآیند رونویسی اطلاعات ژنتیکی DNA و تشکیل rRNA به‌منظور تولید پروتئین، نقش تعیین‌کننده‌ای دارد و یا کمپلکس ریز منفذ هسته ، مسئول مبادله محافظت‌شده اجزاء سلولی میان هسته و سیتوپلاسم پیرامون آن می‌باشد و از انتقال مواد غیرمجاز به درون غشاء هسته جلوگیری می‌کند[۳].

به‌طورکلی با توجه به اهمیت و جایگاه کمپلکس‌های پروتئینی، کشف آن‌ها می‌تواند به درک عمیق‌تر سازوکارهای تنظیمی سلول، توصیف فرآیند تکامل در سیگنال‌های سلولی، پیش‌بینی عملکرد زیستی پروتئین‌های کشف شده و از همه مهم‌تر تحقق اهداف درمانی(تشخیص بیماری و طراحی دارو) منجر شود[٤].

روش‌های آزمایشگاهی برای تشخیص کمپلکس‌های پروتئینی دارای محدودیت‌های مختلفی می‌باشند. برای مثال روش TAP که به کمک طیف‌سنج جرمی انجام می‌شود، مستلزم چندین مرحله شستشو و تصفیه است که منجر به حذف نامطلوب جاذبه(ضعیف و ناپایدار) میان پروتئین‌ها در کمپلکس می‌شود. علاوه بر آن پروتئین‌هایی که به‌عنوان برچسب در این نوع آزمایش‌های تجربی استفاده می‌شوند، می‌توانند در یک اثر نامطلوب، مانع تشکیل کمپلکس‌های پروتئینی شوند. بنابراین، استفاده از روش‌های آزمایشگاهی برای تشخیص کمپلکس‌های پروتئینی بسیار پرهزینه و زمان‌بر می‌باشد و درنتیجه رویکردهای محاسباتی در تشخیص کمپلکس‌ها اهمیت فوق‌العاده‌ای می‌یابد[٥].

در سال‌های اخیر به‌کارگیری فنّاوری‌های پیشرفته آزمایشگاهی در تعیین برهم‌کنش میان پروتئین‌ها منجر به ایجاد حجم بسیار بالایی از داده‌های تجربی شده است که کارکرد پروتئین‌ها را در شبکه‌های سلولی پیچیده منعکس می‌کند. شبکه‌های برهم‌کنش پروتئین-پروتئین منبع داده‌ای اصلی به منظور کشف کمپلکس‌های پروتئینی می‌باشد. انگیزه اصلی از آنجا ناشی شده است که پژوهشگران به تجربه دریافته‌اند که پروتئین‌ها در سلول و در قالب پیمانه‌ها و کمپلکس‌های پروتئینی مختلف وظایف خود را انجام می‌دهند و پروتئین‌هایی که با یکدیگر برهم‌کنش دارند، اغلب در فرآیندهای زیستی یکسانی شرکت می‌کنند و درنتیجه می‌توان در اغلب موارد پیمانه‌های (کمپلکس‌های) پروتئینی را به یک عملکرد زیستی مشخص نگاشت کرد به‌طوری‌که پروتئین‌های یک پیمانه، قرابت عملکردی بیشتری با یکدیگر نسبت به پروتئین‌های یک پیمانه دیگر دارند [٦].

مسئله تشخیص کمپلکس‌های پروتئینی از درون شبکه‌های PPI ازلحاظ محاسباتی به مسئلهٔ خوشه‌بندی این شبکه‌ها نگاشت می‌شود [۷] . انگیزه این کار نیز بر اساس مشاهدات پژوهشگران تجربی از تناظر میان یک کمپلکس واقعی با یک خوشه از پروتئین‌ها در یک شبکه PPI می‌باشد. در یک تقسیم‌بندی، چالش‌های پژوهش در تشخیص کمپلکس‌های پروتئینی را می‌توان به دو دسته کلی چالش‌های ناشی از مجموعه شبکه‌های PPI و چالش‌های روش‌های محاسباتی موجود، تقسیم‌بندی کرد[۸].

شبکه‌های PPI به‌دست‌آمده از روش‌های آزمایشگاهی دارای مقدار زیادی خطای مثبت کاذب می‌باشند و درنتیجه شبکه PPI مورد استفاده خوشه‌بندی، دارای مقدار زیادی نوفه می‌باشد [۹]. با توجه به هزینه و زمان مورد نیاز در انجام روش‌های آزمایشگاهی، استفاده از روش‌های محاسباتی جهت ارتقای کیفیت شبکه‌های برهم‌کنش از طریق وزن‌دار کردن برهم‌کنش‌ها ضروری است. بدیهی است که افزایش اطمینان در منابع داده‌ای موجود به افزایش دقت روش‌های محاسباتی موجود در کاربردهای مختلف منجر می‌شود. متداول‌ترین راهکار جهت بهره‌گیری بیشتر از منابع زیستی و کاهش نوفه (برای مثال، تداخل موجود در شبکه‌های PPI تولیدشده توسط روش‌های آزمایشگاهی

طیف‌سنج جرمی، Yeast2Hybrid و in-silico، وزن‌دار کردن برهم‌کنش‌ها با استفاده از منابع داده‌ای متمایز و با قالب‌های مختلف(مانند پایگاه داده‌های دنباله‌ای، ساختاری، بیان ژن و انواع شبکه‌های با ساختارهای مختلف) می‌باشد[۱۰].

روش‌های تشخیص کمپلکس را می‌توان از دیدهای متفاوتی دسته‌بندی کرد. در یک دسته‌بندی کلان، می‌توان روش‌های تشخیص را برحسب توانایی یا عدم توانایی آن‌ها در مدل‌سازی شرایط پویای سلول در فرآیند تشخیص، به دو دسته روش‌های پویا و روش‌های ایستا تقسیم‌بندی نمود. روش‌های ایستای تشخیص کمپلکس را می‌توان برحسب استفاده و یا عدم استفاده آن‌ها از سایر منابع داده‌ای زیستی به دو دسته تقسیم‌بندی کرد. روش‌های مبتنی بر فقط خوشه‌بندی گراف و روش‌های مبتنی بر خوشه‌بندی گراف همراه با بعضی از دانش‌های اضافی زیست‌شناسی. دانش‌های زیست‌شناسی نیز در قالب قالب هسته–افزونه، شباهت‌های عملکردی، پایسته‌های تکاملی و برهم‌کنش‌های متقابل (یا انحصاری) دسته‌بندی می‌شوند. از مهم‌ترین روش‌هایی که فقط از خوشه‌بندی گراف استفاده می‌کنند می‌توان از روش‌های MCOD[1]،DPClus[11]،CFinder[12]،SPICi[13] CMC[8]، MCL[14]، و ClusterONE[15] را نام برد. از مهم‌ترین روش‌هایی نیز که از اطلاعات زیستی در کنار اطلاعات مربوط به پیکربندی شبکه برهم‌کنش در استخراج کمپلکس‌های پروتئینی استفاده می‌کنند می‌توان به [16] HSS[19]، RNSC[18]، CORE[17]، COACH [16] و [20] TINCD اشاره کرد. از مهم‌ترین روش‌های پویای تشخیص کمپلکس نیز می‌توان به روش‌های TS-[22]، PCD[21]-TSN-OCD و PCD-GED [23] اشاره کرد.

در یک دسته‌بندی دیگر، روش‌های تشخیص بر اساس رویکرد جستجوی کمپلکس‌ها، به چهار حالت مبتنی بر ادغام و توسعه خوشه‌ها ، مبتنی بر افراز شبکه مبتنی بر انطباق شبکه و روش‌های فرامکاشفه‌ای تقسیم می‌شوند. در روش‌های مبتنی بر ادغام و توسعه، کمپلکس‌ها در یک رویکرد پایین به بالا، با شروع از تعدادی خوشه پایه(که می‌تواند یک مثلث و یا کلیک باشد) ایجاد می‌شوند. در روش‌های مبتنی بر افراز شبکه، در یک رویکرد بالا به پایین، هر شبکه به تعدادی زیرشبکه تقسیم می‌شود و در روش‌های مبتنی بر انطباق شبکه نیز از طریق انطباق شبکه‌های چند ارگانیسم زیستی با یکدیگر و استخراج مناطق مشترک، به کشف کمپلکس‌های جدید پرداخته می‌شود. روش‌های مبتنی بر ادغام و توسعه خوشه‌ها نیز به نوبه خود به دو دسته جستجوی چگالی همسایگی محلی LD و مبتنی بر معیارهای آماری SM تقسیم‌بندی می‌شوند. در روش‌های LD، هدف پیدا کردن زیرگراف‌های چگال با یک جستجوی محلی در مجاورت هر خوشه اولیه می‌باشد و روش‌های SM، از معیارهای آماری برای ترکیب خوشه‌ها استفاده می‌کنند. روش‌های مبتنی بر افراز نیز به دو دسته جستجوی محلی مبتنی بر هزینه CL و شبیه‌سازی جریان FS تقسیم‌بندی می‌شوند. در روش‌های CL به افراز شبکه بر اساس بیشینه‌کردن یک تابع هزینه پرداخته می‌شود و در روش‌های FS از مفهوم گام زدن تصادفی برای افراز شبکه استفاده می‌شود[۲۳, ۲٤].

در یک جمع‌بندی درباره نقاط ضعف و قوت رویکردهای جستجوی مختلف باید یادآور شد که روش‌های مبتنی بر ادغام و توسعه از این حیث که امکان کشف خوشه(کمپلکس)های همپوشان را دارا می‌باشند، نسبت به روش‌های مبتنی‌بر افراز که تنها خوشه‌های مجزا را کشف می‌کنند، مزیت دارند. هم‌چنین کارایی روش‌های مبتنی بر انطباق، با توجه به محدودیت و ناقص بودن دانش در دسترس از ارگانیسم‌های مختلف عملاً با محدودیت‌های بسیاری مواجه می‌باشد. روش‌های فرامکاشفه‌ای نیز در تشخیص کمپلکس‌های پروتئینی نتوانسته‌اند به مزیت قابل‌توجهی نسبت به سایر روش‌ها دست یابند. هرچند امروزه، روش‌های تشخیص متنوعی در رویکرد مبتنی بر ادغام و توسعه و در هر دو دسته LD و SM ارائه‌شده

اســت که در زمره بهترین روش‌های تشخیص می‌باشند، اما هم‌چنان چالش اصلی در یک روش تشخیص، چگونگی اســتفاده از دیگر داده‌های زیستی به منظور مقابله با نوفهٔ موجود در شــبکه‌های PPI و در نظرگرفتن شرایط پویای سلول در تشخیص کمپلکس‌ها می‌باشد[۲۵].

۲. مفاهیم اولیه

در این بخش، مروری بر مفاهیم اولیه‌ای که برای درک بهتر روش ارائه شده ضروری است انجام می‌گیرد.

شبکه برهم‌کنش پروتئین- پروتئین

به‌طورکلی اکثر فرآیندهای درون‌ســلولی مانند انتقال سیگنال‌های الکتروشیمیایی، نقل‌وانتقال مواد، تنظیم موازنه یونی درون هسته و غیره، مستلزم تشکیل پیوند(برهم‌کنش) میان محصــولات ژنی(اعم از پروتئین‌هــا، RNA و غیره) می‌باشد. در سال‌های اخیر توسعه روش‌های آزمایشگاهی پیشــرفته با بازده بالا (از قبیــل ژل الکتروفورز دو بعدی و طیف‌ســنج جرمی) در تعیین برهم‌کنش میان پروتئین‌ها منجر به‌ایجاد حجم بســیار بالایی از داده‌های تجربی شده است که کارکرد پروتئین‌ها را در شبکه‌های سلولی پیچیده منعکس می‌کند. شــبکه‌های برهم‌کنــش پروتئین-پروتئین یکی از منابع داده‌ای اصلی در زیست‌شناســی ســامانه‌ها و مهم‌ترین منبــع داده‌ای به منظور کشــف کمپلکس‌های پروتئینی می‌باشــند. در این شبکه‌ها، رئوس گراف متناظر بــا پروتئین‌ها و یال‌ها نتیجه یک نوع برهم‌کنش میان آن‌ها می‌باشند.

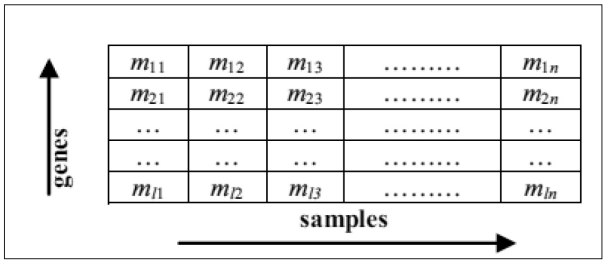
کمپلکس پروتئینی

یــک کمپلکس پروتئینی به منزله یک ماشــین مولکولی اســت که شــامل تعدادی پروتئین می‌باشد که در یک بازه زمانی و مکانی مشــخص به یکدیگر متصل می‌شوند. اکثر پروتئین‌ها عملکردهای زیســتی خود را در سلول زنده در قالب تشکیل کمپلکس‌ها انجام می‌دهند. در حقیقت بسیاری از پروتئین‌هــا به عنوان اجــزای کمپلکس‌های ضروری و پایدار وجود دارند. برای مثال مولکول حیاتی هموگلوبین، یک ترکیب پایــدار از چهار پروتئین گلبول دار مشــخص می‌باشــد. هم‌چنین بســیاری از آنزیم‌ها که نقش راه‌انداز را در بســیاری از واکنش‌های درون سلول به عهده‌دارند، مجموعه‌ای از پروتئین‌ها می‌باشــند که به‌طور هم‌زمان به یکدیگر متصل شده و یک آنزیم را شکل می‌دهند. علاوه بر آن برهم‌کنش‌های ناپایدار و موقت مانند اثرات هورمون و یا اثرات یک ســیگنال موقت، مستلزم تشکیل کمپلکس‌های پروتئینی می‌باشد. از لحاظ زیستی، کمپلکس‌های پروتئینی اجزاء کلیدی مولکولی برای انجام بسیاری از عملکردهای زیســتی از قبیل رونوشــت DNA، ترجمــه mRNA، انتقال سیگنال سلولی، چرخه سلولی و غیره می‌باشد.

بیان ژن و ماتریس آن

ریزآرایه دی‌ان‌ای[1]، یک ســطح جامد مشبک متشکل از هزاران خانه می‌باشد، به‌طوری‌که هر خانه در سطح شبکه، متناظر با یک ژن می‌باشــد. به منظــور انجام یک آزمایش ریزآرایه، دو نمونه سالم(یا کنترل) و نمونه هدف(که معمولاً یک mRNA از قبیل عامل رونویســی یا یک عامل بیماری و یا هر محصول ژنی موردمطالعه می‌باشــد، تهیه می‌شود. ســپس، به هر یک از دو نمونه، یک رنگ به‌عنوان برچسب نظیر می‌شــود. در مرحله بعدی، با مخلوط کردن محتوای دو نمونه، مقدار متناسبی از آن به صورت ترکیبات مایع بر روی ســطح ریزآرایه ریخته می‌شود و در نهایت برحسب شدت و ضعف پیوند (با توجه به تمایل ترکیبات نوکلئوتیدی به اتصال به مکمل خود، این پیوند برقرار می‌شــود) میان نمونــه با هر یک از ژن‌های ســطح ریزآرایــه، به ازای هر ژن در هــر آزمایش و به کمک روش‌های پردازش تصویر یک کمیّت عددی به دســت می‌آید که بیانگر مقدار بیان ژن در آزمایش مربوطه اســت. به طور خلاصه می‌توان گفت که با یــک آزمایش ریزآرایه امکان ثبت مقدار بیان هزاران ژن فراهم می‌شــود. در هنگام استفاده از ریزآرایه، معمولاً تعــدادی آزمایش متوالی به‌صورت یک ســری از فواصل

1- DNA Micro array



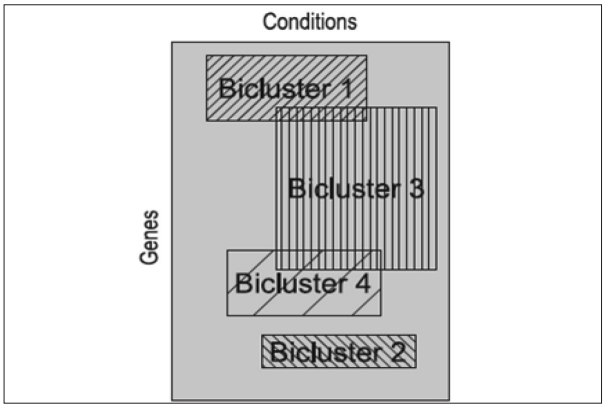
شکل ۱: ماتریس بیان ژن[۲۶]

زمانی یکسان درون چرخه سلولی و یا تعدادی آزمایش در شرایط آزمایشگاهی متفاوت (برای مثال بافت‌های مختلف و یا آزمایش در حضور محصولات ژنتیکی مختلف) انجام می‌گیرد. خروجی n آزمایش یک ریزآرایه شامل l ژن، یک ماتریس $M_{(l\times n)}$ به نام ماتریس بیان ژن می‌باشد که در آن m_{ij} مقدار بیان‌ژن iام را در آزمایش jام نشان می‌دهد(شکل ۱). به سطر متناظر با مجموعه مقادیر بیان یک ژن در ماتریس، پروفایل بیان آن ژن می‌گویند و ژن‌های دارای پروفایل بیان شبیه به هم را هم-بیان[۲] و ژن‌های دارای عملکرد یکسان در یک یا تعدادی از شرایط را هم-تنظیم[۳] می‌گویند.

خوشه‌بندی دوگانه[۴]

فرض کنید یک مجموعه داده شامل n نمونه داده باشد که هر نمونه m خصیصه دارد. در الگوریتم‌های خوشه‌بندی، محاسبه فاصله دو نمونه داده، فاصله بین تمامی m خصیصه در نظر گرفته می‌شود. هر عنصر داده‌ای را در تعیین فاصله عنصر داده‌ای با بقیه عناصر لحاظ می‌کند و این در حالی است که دو عنصر داده‌ای می‌توانند فقط در زیرمجموعه‌ای از خصیصه‌ها به یکدیگر شبیه باشند. خوشه‌بندی دوگانه که به آن خوشه‌بندی دوبعدی نیز می‌گویند، یکی از روش‌های یادگیری ماشین است که عمل خوشه‌بندی را به طور هم‌زمان بر روی سطرها (نمونه‌ها) و ستون‌های(خصیصه‌های) ماتریس داده انجام می‌دهد. خروجی یک الگوریتم خوشه‌بندی دوگانه، تعدادی خوشه دوگانه از نمونه‌ها می‌باشد به‌طوری‌که هر خوشه دوگانه،

2- Co-express Genes
3- Co-regulated Genes
4- BiClustering



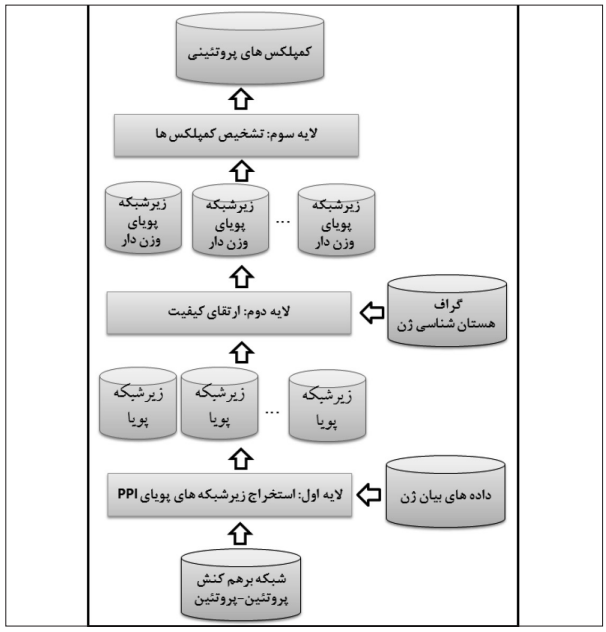
شکل ۲: خوشه‌بندی دوگانه ماتریس بیان ژن[۲۷]

حاوی تعدادی نمونه داده است که در زیرمجموعه‌ای از خصیصه‌ها دارای بیشترین شباهت به یکدیگر می‌باشند. شکل ۲ خوشه‌بندی دوگانه ماتریس بیان ژن، جهت استخراج مجموعه ژن‌های با الگوی بیان به هم-وابسته را نشان می‌دهد.

گراف هستان‌شناسی ژن[۵]

توسعه فنّاوری منجر به تولید داده‌ها و دانش زیادی در هریک از زمینه‌های پژوهشی شده است و این مسئله ضرورت استانداردسازی و سازمان‌دهی این حجم بالای اطلاعات را به‌منظور تسهیل دسترسی و استخراج دانش نهفته در آن‌ها اجتناب‌ناپذیر ساخته است. با استفاده از گراف مفهومی هستان‌شناسی، مدل‌سازی دانش کشف‌شده در قالب تعدادی مفاهیم(ترم اطلاعاتی) و رابطه معنایی میان آن‌ها امکان‌پذیر است. در یک گراف هستان‌شناسی، به طور عمده، دو رابطه is-a و part-of میان ترم‌های اطلاعاتی وجود دارد. رابطه is-a، به معنای یک نمونه خاص از یک مفهوم(مانند درخت سیب، یک نمونه از درخت است) و رابطه part-of، به معنای بخشی از یک مفهوم دیگر است(مانند منقار و پر در عقاب). یک گراف هستان‌شناسی ژن یا به اختصار GO، یک ساختار سلسله مراتبی درخت‌واره از ترم‌های زیستی است که رابطه is-a و part-of این ترم‌ها را نشان می‌دهد. در یک نگاه دقیق‌تر، گراف هستان‌شناسی ژن از سه گراف هستان‌شناسی بدون دور متناظر با مؤلفه

5- Gene Ontology Graph (GO graph)

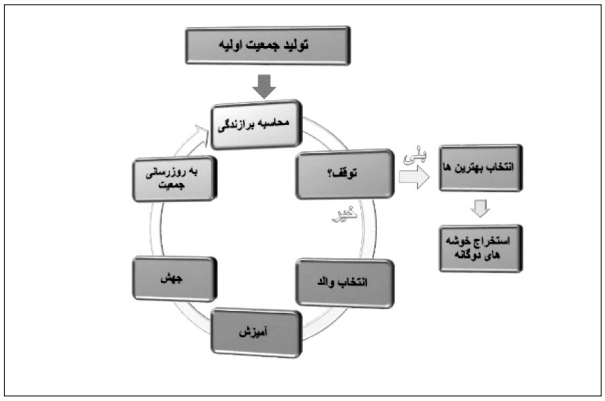


شکل ۳: مدل سه لایه پیشنهادی

سلولی یا به‌اختصار CC، عملکرد مولکولی یا به اختصار MF و فرآیند زیستی یا به‌اختصار BP تشکیل‌شده است. هر ترم زیستی در GO، شامل تعدادی محصول ژنی از قبیل پروتئین است. برای مثال، یک ترم زیستی می‌تواند بیانگر یک فرآیند زیستی باشد که از چندین محصول ژنی تشکیل‌شده است. از این حیث، می‌توان گفت که هر ترم، تعدادی ژن را تفسیر می‌کند. این تفسیر، می‌تواند مستقیم و یا غیرمستقیم و از طریق رابطه سلسله مراتبی درخت باشد و معمولاً پایه‌ای برای محاسبه مقدار اطلاعات فراهم‌شده توسط ترم می‌باشد.

۳. مواد و روش‌ها

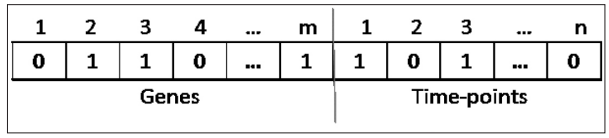
شکل ۳ مدل پیشنهادی مقاله را در قالب سه لایه نشان می‌دهد. در ادامه، به معرفی مختصر هر یک از لایه‌های مدل پرداخته می‌شود. با توجه به این که در نظر گرفتن پویایی سلول در تشخیص کمپلکس‌های پروتئینی، یکی از چالش‌های اصلی مسئله تشخیص می‌باشد. لایه اول، با استفاده از داده‌های بیان ژن، راهکاری را به این منظور پیشنهاد می‌دهد. ورودی لایه اول، شبکه PPI و یک ماتریس بیان‌ژن می‌باشد و خروجی این لایه، تعدادی زیرشبکه PPI



شکل ۴: مراحل اصلی الگوریتم GA-GCT

می‌باشد. نحوه کار بدین صورت است که ابتدا، توسط الگوریتم ژنتیک پیشنهادی GA-DCT، به خوشه‌بندی دوگانه داده‌های بیان ژن پرداخته می‌شود. در این الگوریتم (شکل ٤) به منظور استخراج خوشه‌های دوگانه از ماتریس بیان ژن استفاده می‌شود. به‌طور خلاصه، در مقایسه با سایر روش‌های خوشه‌بندی دوگانه، این الگوریتم دارای دو ویژگی متمایز می‌باشد: ۱) مبتنی بر روش فرامکاشفه‌ای الگوریتم ژنتیک و در نتیجه جستجوی خوب در فضای راه‌حل‌های مسئله ۲) مبتنی بر گسسته‌سازی ماتریس بیان ژن و در نتیجه مقاوم به نوفه موجود در این داده‌ها. نحوه کار بدین صورت است که ماتریس بیان ژن M با m سطر و n ستون داده شده است هر عنصر m_{ij} از ماتریس M، مقدار بیان i امین ژن را در j امین نقطه زمانی نشان می‌دهد. یک خوشه دوگانه B(I,J) یک زیر ماتریس از M می‌باشد. چنین خوشه دوگانه‌ای، به کمک یک رشته دودویی (به‌عنوان کروموزوم) شامل دو قسمت نشان داده می‌شود. قسمت اول کروموزوم به طول m، متناظر با نمایه‌های ژن‌های موجود در خوشه دوگانه و قسمت دوم به طول n، متناظر با نمایه‌های نقاط زمانی موجود در خوشه دوگانه می‌باشد. بنابراین، هر خوشه دوگانه به وسیله یک بردار دودویی با طول ثابت n+m که برحسب سطرها و ستون‌های موجود در خوشه، تعدادی از بیت‌های آن در دو قسمت دارای مقدار ۱ و مابقی صفر می‌باشد(شکل ٥).

پس از حذف خوشه‌های کم‌اهمیت، متناظر با هر خوشه دوگانه، یک زیرشبکه PPI از شبکه اولیه PPI استخراج



شکل ۵: بیان یک خوشه دوگانه در یک کروموزوم

می‌شود به‌طوری‌که هر زیرشبکه، به شامل برهم‌کنش‌های بین پروتئینی ایجادشده در بازه‌ای از زمان می‌باشد و در نتیجه می‌توان آن را به عنوان یک زیرشبکه پویا نام‌گذاری کرد. مجموعه زیرشبکه‌های پویای PPI استخراج شده، ورودی لایه دوم مدل می‌باشد. یکی از چالش‌های مهم مسئله، «وجود نوفه در شبکه PPI می‌باشد [۲۸] و از طرفی منابع اطلاعاتی دیگری مانند گراف هستان‌شناسی ژن، وجود دارد که استفاده از آن‌ها در کنار شبکه‌های PPI می‌تواند به کاهش نوفه و افزایش دقت این شبکه‌ها منجر شود. هدف اصلی لایه دوم، ارتقای کیفیت شبکه‌های PPI موجود از طریق وزن‌دار کردن برهمکنش‌های بین پروتئینی می‌باشد. منظور از وزن نیز یک کمیت عددی بین صفر و یک می‌باشد که بیانگر قوت و یا ضعف اتصال میان دو پروتئین می‌باشد. در این لایه، هر زیرشبکه از مجموعه زیرشبکه‌های پویای PPI به‌دست‌آمده در لایه قبلی، به صورت جداگانه وزن‌دار می‌شود. این عمل باعث می‌شود که مبنای پویایی به‌دست‌آمده در لایه قبل، در این لایه نیز لحاظ گردد و در واقع، عمل وزن‌دار کردن زیرشبکه‌های پویای PPI به‌صورت محلی انجام می‌گیرد. وزن محاسبه شده به ازای هر زوج پروتئین در دو سر یک یال(برهم‌کنش) با روش پیشنهادی SSW تعیین می‌شود که ترکیبی از دو مقدار شباهت دو پروتئین متناسب با دو منبع داده‌ای مختلف می‌باشد. این منابع داده‌ای عبارتند از شبکه PPI و گراف هستان‌شناسی ژن. شبکه PPI، اطلاعات پیکربندی (برهم‌کنش‌های فیزیکی) لازم برای اندازه‌گیری شباهت را منعکس می‌کند و گراف هستان‌شناسی ژن، دربرگیرنده رابطه و جایگاه عملکردی پروتئین‌ها در کل ژنوم و در یک ساختار سلسله مراتبی می‌باشد. در این لایه، برای محاسبه شباهت مبتنی بر گراف هستان‌شناسی ژن، یک سنجنده شباهت معنایی ترکیبی(مبتنی‌بر گره و مبتنی‌بر یال) دو مرحله‌ای به نام Tsss ارائه شده است که توانسته است نسبت به سنجنده‌های متعارف معنایی به دقت بهتری دست یابد.

تابع SSW(x,y)، وزن متناظر با برهم‌کنش میان پروتئین‌های x و y را با استفاده از مقادیر شباهت مبتنی بر پیکربندی و شباهت معنایی این دو پروتئین محاسبه می‌کند. براساس رابطه (۱)، SSW(x,y)، ترکیبی وزن‌دار از دو تابع Tsss(x,y) و Sim_{top} (x,y) می‌باشد که α وزن هریک از این دو نوع شباهت در محاسبه SSW می‌باشد.

$$SSW(x,y)=\alpha\times Sim_{topo}(x,y)+(1-\alpha)\times Tsss(x,y) \quad (۱)$$

در این رابطه، Sim_{top} (x,y)، می‌تواند هریک از توابع محاسبه مقدار شباهت مبتنی بر پیکربندی شبکه PPI باشد. بدیهی است که با انتخاب مقدار مناسب برای α، می‌توان تأثیر هر یک از دو مقدار شباهت مبتنی بر پیکربندی شبکه و گراف هستان‌شناسی را کنترل کرد. تابع Tsss(x,y)، مقدار شباهت معنایی پروتئین‌های x و y را در دو مرحله محاسبه می‌کند. در مرحله اول ماتریس شباهت‌های ترم-ترم SS را برای تمامی ترم‌های موجود در گراف GO محاسبه می‌کند و در مرحله دوم، با استفاده از ماتریس SS، شباهت‌های معنایی میان پروتئین‌ها محاسبه می‌کند. در مرحله اول، شباهت معنایی میان دو ترم x و y در GO بر اساس رابطه (۲) محاسبه می‌شود که در آن C_t و C_g دو ضریب شباهت معنایی مبتنی بر ترم می‌باشند. رابطه C_t (x,y)، تابعی از اجداد مشترک دو ترم x و y می‌باشد و براساس (۳) محاسبه می‌شود. منظور از A_x و A_y، مجموعه اجداد ترم‌های x و y و منظور از IIC(a) نسخه معکوس شده IC (a)، تابع محتوای اطلاعات[6] بر اساس رابطه (٤) برای ترم a می‌باشد.

$$SS(x,y) = C_t(x,y) \times C_g(x,y) \quad (۲)$$

$$C_t(x,y)= \sum_{a\in A_x\cap A_y} \frac{2\times IIC(a)}{SV(x)+SV(y)} \quad (۳)$$

$$IIC(a) = 1-IC(a) \quad (٤)$$

6 Information Content (IC)

مقدار SV(a)، نیز مقدار معنایی ترم a، متناظر با جمع معنایی مقادیر اطلاعات معکوس شده در رابطه (۵) می‌باشد. پارامتر W_t نیز با استفاده از رابطه (٦) محاسبه می‌شود.

$$SV(a) = \sum_{t \in Aa} w_t \times IIC(t) \quad (۵)$$

$$W_t = Min(\text{distance t to GO root}) / Min(\text{length of path from GO root to a leaf node cross t}) \quad (٦)$$

مطابق روابط اخیر، ضریب $C_t(x,y)$ در رابطه (۳)، ترکیبی از رویکرد مبتنی بر گره و مبتنی بر یال می‌باشد. مقدار $C_g(x,y)$ نیز، به عنوان دومین ضریب شباهت معنایی در یک رویکرد مبتنی بر مجموعه، بر اساس رابطه (۷) محاسبه می‌شود که D_x و D_y مجموعه نواده‌های ترم‌های x و y می‌باشد. بنابراین، روش Tsss در مرحله اول، ترکیبی از رویکردهای مبتنی بر گره، مبتنی بر یال و مبتنی بر مجموعه را برای اندازه‌گیری شباهت معنایی استفاده می‌کند. یادآوری می‌گردد این مرحله فقط یک بار پس از دریافت گراف هستان‌شناسی ژن انجام می‌شود.

$$C_g(x,y) = |D_x \cap D_y| / |D_x \cup D_y| \quad (۷)$$

در مرحله دوم، شباهت معنایی میان پروتئین‌های a و b $Tsss_t(a,b)$، بر اساس رابطه (۸) و با استفاده از ماتریس SS حاصل از مرحله اول محاسبه می‌شود.

$$Tsss_t(a,b) = ss_t(a,b) \times g(a,b) \quad t \in \{MAX, AVG, BMA\} \quad (۸)$$

sst (a,b)، با رویکردی مبتنی بر زوج، شباهت میان پروتئین‌های a و b را برحسب شباهت‌های ترم-ترم در مجموعه ترم‌های تفسیرکننده آن‌ها و به کمک یکی از قوانین AVG، MAX و یا BMA محاسبه می‌کند. g (a,b) نیز یک سنجنده مبتنی‌بر گراف است که با استفاده از رابطه (۹) محاسبه می‌شود.

$$g(a,b) = \frac{\sum_{t \in f_\theta(N_a \cap N_b)} IC(t)}{\sum_{t \in f_\theta(N_a \cup N_b)} IC(t)} \quad (۹)$$

که N_a و N_b مجموعه ترم‌های تفسیر پروتئین‌های a و b و f_θ، یک تابع پالایه می‌باشد که برحسب یک مقدار آستانه θ، ترم‌های با عمق کم (نزدیک به ریشه درخت GO) را از مجموعه از ترم‌ها حذف می‌کند تا از تولید مقادیر شباهت

CAMWI Method

Input: G(V, E): A weighted PPI network, α, β : seed and set thresholds
Output: Complexes: the detected protein complexes
1-Seeds ← SeedSelection(G, α)
2-Cores ← CoreFinding(G,Seeds)
3-InitComplexes ← CoreGrowing(G,Cores, β)
4-Complexes ← Filtering(InitComplexes)

شکل ۶: مراحل مختلف روش CAMWI

زیاد و نزدیک به هم برای اکثر زوج‌های پروتئینی جلوگیری شود.

لایه سوم، به هدف اصلی مدل یعنی تشخیص کمپلکس‌های پروتئینی از درون مجموعه زیرشبکه‌های پویای وزن‌دار PPI به‌دست‌آمده از لایه دوم می‌پردازد. روش ارائه‌شده با نام CAMWI، کمپلکس‌ها را در یک قالب هسته-افزونه تشخیص می‌دهد و ذاتاً برای شبکه‌های وزن‌دار ارائه شده است هرچند که استفاده از آن در شبکه‌های بدون وزن نیز برتری آن را نسبت به سایر روش‌ها نشان می‌دهد. چهار قدم اساسی این روش عبارت است از ۱) انتخاب دانه‌های اولیه به کمک مفهوم ضریب خوشه‌بندی وزن‌دار ۲) شناسایی هسته پیرامون هر دانه با استفاده از مفهوم چگالی وزن‌دار در یک رویه حریصانه، به‌طوری‌که هر هسته، دارای چگالی وزن‌دار بیشینه می‌باشد ۳) توسعه هر هسته کمپلکس با اضافه‌کردن پروتئین‌های افزونه و تشکیل کمپلکس ٤) پالایش و حذف کمپلکس‌های تکراری و یا بسیار شبیه به هم. شکل ٦ الگوریتم CAMWI را نشان می‌دهد.

۴. نتایج

در این بخش، نتایج ارزیابی مدل پیشنهادی که با تجمیع لایه‌های سه‌گانه به دست می‌آید، ارائه می‌شود. مدل پیشنهادی، در لایه اول، به کمک روش پیشنهادی GA-DCT، به پویاسازی شبکه PPI می‌پردازد، به‌گونه‌ای‌که لایه اول، با دریافت شبکه PPI ورودی، زیرشبکه‌های پویای PPI را بر مبنای داده‌های بیان ژن تولید می‌کند. سپس، مجموعه

جدول ۱: منابع داده ای مورد استفاده

توصیف	نوع	مجموعه داده
تعداد ۵۹۷۴۸ برهم‌کنش میان ۵۶۴۰ پروتئین	شبکه برهم‌کنش	BioGrid [۲۹]
سطح بیان ۳۵۵۲ ژن در ۱۲ آزمایش	ماتریس بیان ژن	YMC [۳۰]

زیرشبکه‌های پویای PPI، جهت ارتقای کیفیت در اختیار لایه دوم قرار می‌گیرد. لایه دوم، با استفاده از روش پیشنهادی SSW، به وزن‌دار کردن مستقل زیرشبکه‌های پویا می‌پردازد و مجموعه زیرشبکه‌های پویای وزن‌دار PPI در اختیار لایه سوم که وظیفه اصلی تشخیص کمپلکس‌ها را برعهده دارد، قرار می‌گیرد. لایه سوم، با استفاده از روش پیشنهادی CAM-WI، کمپلکس‌های پروتئینی را در هر زیرشبکه پویای وزن‌دار تشخیص می دهد. مجموعه کمپلکس‌های به دست‌آمده به ازای تمامی زیرشبکه‌ها، پس از پالایش و حذف کمپلکس‌های تکراری و یا بسیار شبیه به‌هم، مجموعه کمپلکس‌های نهایی تشخیص‌داده شده توسط مدل پیشنهادی می‌باشد.

به منظور ارزیابی روش محاسباتی ارائه شده از سه نوع مجموعه داده مختلف استفاده شده است: یک شبکه برهم کنش پروتئین-پروتئین، یک مجموعه داده محک از کمپلکس‌های پروتئینی کشف شده به روش آزمایشگاهی و یک مجموعه داده بیان ژن. جدول ۱ و ۲ ویژگی منابع داده ای مورد استفاده در این ارزیابی را نشان می‌دهد.

از آنجا که در تحلیل‌های ارائه شده و در مقایسه با روش‌های گوناگون، با توجه به شرایط و دستگاه مقایسه، از مجموعه داده‌های متفاوتی استفاده شده است، در این بخش، در یک دستگاه ارزیابی یکسان، هر یک از مراحل توسعه مدل به عنوان یک روش نام‌گذاری می‌شود و با حالت کامل مدل پیشنهادی(شامل سه لایه) مقایسه شد. نتایج مربوطه در شکل ۲ نشان داده شده است. در این مقایسه، از روش‌های زیر استفاده شده‌است:

- فقط لایه سوم، با نام CAMWI.
- لایه دوم، سپس لایه سوم، با نام SSW+CAMWI: در

جدول ۲: مشخصات مجموعه کمپلکس‌های محک مورد استفاده در ارزیابی روش

مجموعه محک	تعداد کمپلکس	تعداد پروتئین	اندازه کمپلکس			
			3 >	10 - 3	11-25	25 <
CYC2008 [31]	408	1627	172	204	27	5

ابتدا، برهم‌کنش‌های بین پروتئینی به روش SSW (لایه‌دوم) وزن‌دار می‌شود، سپس عمل تشخیص به کمک روش CAMWI انجام می‌گیرد.

- لایه اول، سپس لایه سوم، با نام DCT+CAMWI: این روش به نام Bi-CAMWI نام‌گذاری شد. در این روش، ابتدا، با استفاده از روش پیشنهادی GA-DCT، زیرشبکه‌های پویا، از شبکه PPI اولیه، استخراج می‌شود و سپس عمل تشخیص به کمک روش CAMWI بر روی هر زیرشبکه انجام می‌شود.

-لایه دوم، لایه اول و سپس لایه سوم، با نام SSW+DCT+CAMWI: در این روش، در ابتدا، شبکه PPI به کمک روش SSW به صورت سراسری وزن‌دار می‌شود و سپس با استفاده از روش پیشنهادی GA-DCT، زیرشبکه‌های پویای وزن‌دار، از شبکه PPI وزن‌دار، استخراج می‌شود و سپس عمل تشخیص به کمک روش CAMWI بر روی هر زیرشبکه انجام می‌شود.

-لایه اول، لایه دوم و سپس لایه سوم، با نام DCT+SSW+CAMWI: این روش، مدل سه‌لایه پیشنهادی مقاله می‌باشد.

دقت در نتایج به‌دست‌آمده نتایج زیر را نشان می‌دهد.:

۱. مقایسه میان دو روش CAMWI و SSW+CAMWI نشان‌می‌دهد که وزن‌دار کردن شبکه PPI، از این حیث که باعث می‌شود زیربرنامه انتخاب دانه‌ها به صورت هوشمندانه‌تر، زیرمجموعه کوچکتری از دانه‌های اولیه را انتخاب کند؛ منجر به افزایش مقدار سنجنده صحّت می‌شود. از طرف دیگر بدیهی است که کاهش تعداد کمپلکس‌های تشخیصی به کاهش مقدار سنجنده بازخوانی منجر شده است، هرچند که در کل مقدار سنجنده f_measure افزایش

یافته است. این نتایج، تأییدی است بر وجود مقادیر خطای بالای مثبت کاذب در شبکه PPI، به‌گونه‌ای که کاهش نوفهٔ موجود در شبکه PPI از طریق وزن‌دار کردن آن، توانسته است از تولید جواب‌های اشتباه فراوان توسط روش جلوگیری نموده، منجر به افزایش قابل ملاحظه‌ای در مقدار سنجنده صحّت شود. این مسئله در بهبود مقدار سنجنده MMR نیز مشاهده می‌شود که بیانگر افزایش میزان انطباق کمپلکس‌های پیش‌بینی شده با کمپلکس‌های واقعی پس از ارتقای کیفیت شبکه PPI با روش SSW می‌باشد.

۲. مقایسه میان دو روش CAMWI و DCT+CAMWI نشان می‌دهد که در نظرگرفتن شرایط پویای سلول، منجر به بازیابی مجموعه‌ای جدید از کمپلکس‌های پروتئینی شده است و به همین دلیل نرخ سنجنده بازخوانی در DCT+CAMWI افزایش قابل ملاحظه‌ای نسبت به CAMWI پیدا کرده است.

۳. مقایسه میان دو روش SSW+CAMWI و DCT+ CAMWI نشان می‌دهد که تعدادی از کمپلکس‌های پروتئینی وجود دارند که جز از طریق در نظرگرفتن شرایط پویای سلول قابل بازیابی نمی‌باشند؛ به همین دلیل نرخ سنجنده بازخوانی در DCT+CAMWI افزایش قابل ملاحظه‌ای نسبت به SSW+CAMWI پیدا کرده است؛ از طرف دیگر به علت وزن‌دار بودن برهم‌کنش‌ها در SSW+CAMWI با استدلالی مشابه حالت ۱، نرخ صحّت در آن بیشتر می‌باشد.

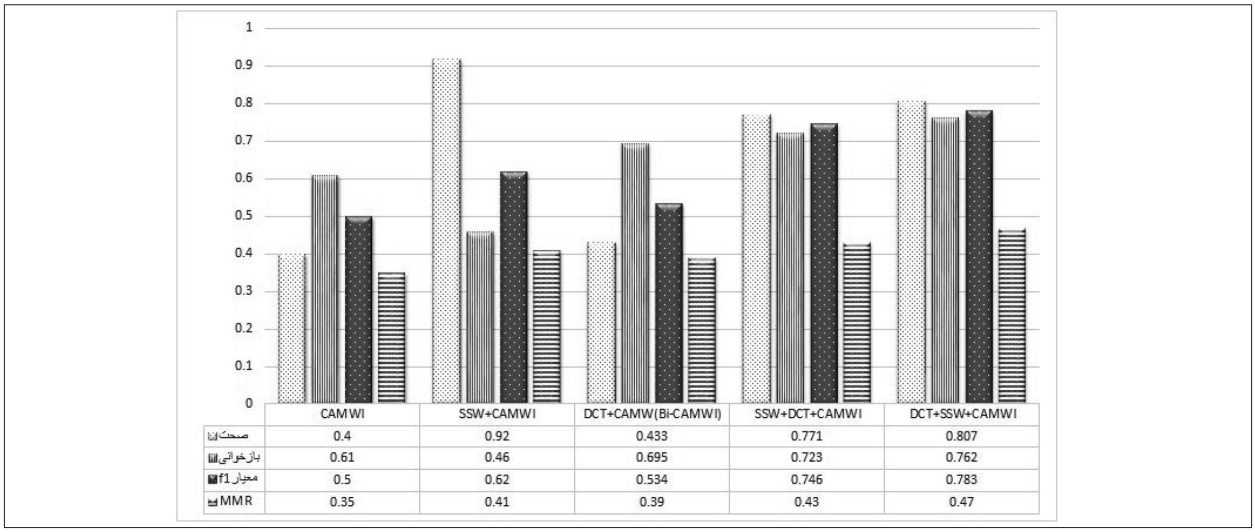
٤. در مقایسه مدل کامل DCT+SSW+CAMWI با روش دوم یعنی SSW+CAMWI، بدیهی است که در نظرگرفتن پویایی از این حیث که توانایی تشخیص کمپلکس‌های پویا را امکان‌پذیر می‌سازد، منجر به افزایش نرخ بازخوانی شده است و از طرف دیگر به علت افزایش تعداد کمپلکس‌های تشخیصی(به علت افزایش تعداد شبکه‌های ورودی) نرخ صحّت کاهش یافته است.

۵. در مقایسه مدل کامل با روش سوم یعنی DCT+CAMWI، بدیهی است که وزن‌دار کردن برهم‌کنش‌ها در مدل کامل، توانسته است نرخ صحّت را افزایش دهد. هم‌چنین، این بار برعکس مقایسه روش اول و دوم، از آنجا که عمل وزن‌دار کردن به‌صورت محلی(برای هر زیرشبکه، به‌صورت مستقل) انجام گرفته است، مقدار بازخوانی نیز افزایش یافته است. درواقع، وزن‌دار کردن محلی، به بازخوانی بهتر کمپلکس‌ها منجر شده است. این مسئله در بهبود مقدار سنجنده MMR نیز مشاهده می‌شود که بیانگر افزایش میزان انطباق کمپلکس‌های پیش‌بینی شده با کمپلکس‌های واقعی در مدل کامل می‌باشد.

٦. مقایسه مدل چهارم SSW+DCT+CAMWI با مدل کامل DCT+SSW+CAMWI نیز نشان می‌دهد که وزن‌دار کردن محلی در مدل کامل، منجر به افزایش مقدار سنجنده صحّت شده است، این مسئله نشان دهنده تاثیر منفی برهم‌کنش های بین پروتئینی نوفه‌دار در حالت وزن‌دار کردن سراسری است. هم‌چنین، افزایش کیفیت اتصالات زیرشبکه پویا در حالت مدل کامل، به بازخوانی بهتر کمپلکس‌ها نیز منجر شده است. این مسئله در بهبود مقدار سنجنده MMR نیز مشاهده می‌شود که بیانگر افزایش میزان انطباق کمپلکس‌های پیش‌بینی شده با کمپلکس‌های واقعی در مدل کامل می‌باشد.

در ادامه، در یک ارزیابی متفاوت، روش‌های تشخیص قبلی که از ترکیبات مختلف لایه‌های سه‌گانه مدل پیشنهادی ساخته شده‌اند، از حیث تعداد کمپلکس‌های تشخیص داده شده در بازه‌های مختلف برحسب اندازه کمپلکس، مقایسه شدند. نتایج مربوطه در جدول ۲ نشان داده شده است. این نتایج، به طور خاص، اثر وزن‌دار کردن شبکه و کاهش نوفهٔ آن به کمک روش SSW را در کاهش تعداد کمپلکس‌های پیش‌بینی شده با اندازه کوچکتر از ۳(کمپلکس‌های نوفه‌دار) نشان می‌دهد.

با توجّه به ماهیت تصادفی مدل سه لایه پیشنهادی، نتایج آن از یک اجرا به اجرای دیگر متفاوت است. بنابراین، به منظور اندازه‌گیری بازه تغییرات نتایج، مدل پیشنهادی در بهترین مقادیر پارامتر به ازای شبکه برهم‌کنش BioGrid و داده‌های بیان ژن YMC و مجموعه محک CYC2008،



شکل ۷: نتایج بررسی مدل پیشنهادی

جدول ۲: مقایسه ترکیبات مختلف لایه‌های سه‌گانه مدل پیشنهادی برحسب تعداد کمپلکس‌های پیش‌بینی شده در اندازه‌های مختلف

روش	تعداد کمپلکس‌های پیش‌بینی شده			
	کمتر از ۳	۳-۱۰	۱۱-۲۵	بیشتر از ۲۵
CAMWI	۱۰۵	۲۱۱	۱۱	۱
SSW+CAMWI	۵۱	۵۷	۵	۲
DCT+CAMWI	۹۶	۱۷۱	۶	۱
SSW+DCT+CAMWI	۷۵	۶۸	۴	۳
DCT+SSW+CAMWI	۶۷	۷۲	۳	۲

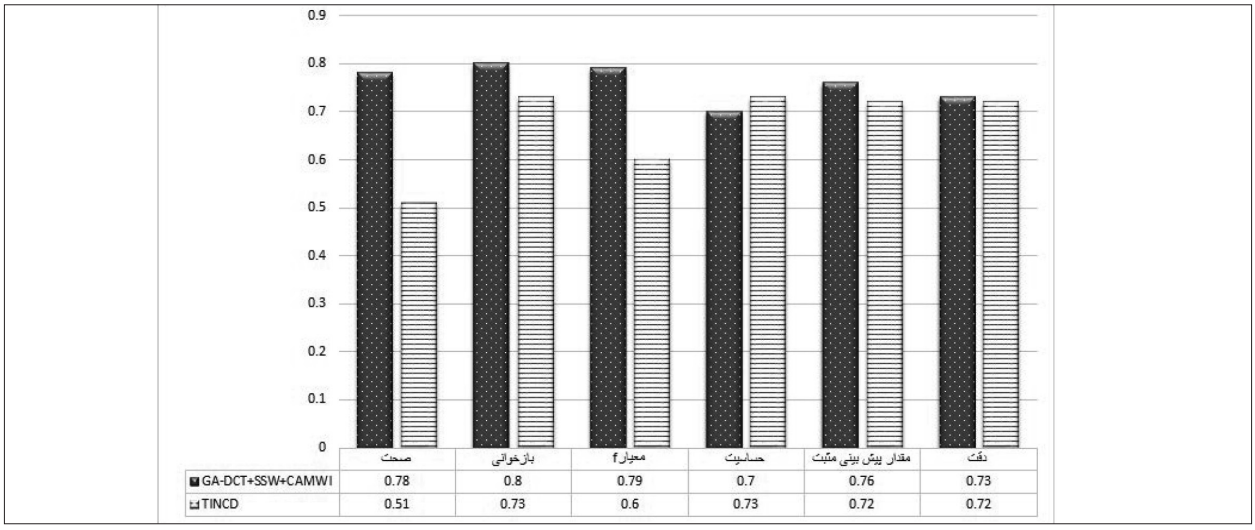
به تعداد ۱۰ بار اجرا شد و مقادیر سنجنده‌های صحّت، بازخوانی و سنجنده اندازه‌گیری شد. محاسبه میانگین و انحراف معیار هر یک از سنجنده‌ها در جدول ۳ نشان داده شده است. این نتایج نشان می‌دهد که برای دستیابی به یک دقّت قابل قبول برای روش، اجرای حداقل ۱۰ تکرار از الگوریتم ضروری است.

در ادامه ارزیابی‌های این بخش، به مقایسه مدل سه لایه پیشنهادی با روش TINCD [۲۰] پرداخته شد. روش TINCD، از این حیث که یکی از جامع‌ترین روش‌هایی است که تاکنون به منظور تشخیص کمپلکس‌های پروتئینی ارائه شده است، یک معیار ارزیابی مناسب می‌باشد. نتایج ارزیابی به ازای سنجنده‌های صحّت، بازخوانی،f1، حسّاسیت، مقدار پیش‌بینی‌شده مثبت و دقّت، در شکل ۳

جدول ۳: مقادیر میانگین و انحراف معیار سنجنده‌های مختلف در ۱۰ اجرای یکسان مدل پیشنهادی با بهترین مقادیر پارامتر

سنجنده	میانگین	واریانس
صحّت	۰,۷۸	۰,۰۰۱۲
بازخوانی	۰,۷۳	۰,۰۰۱۴
f_1	۰,۷۶	۰,۰۰۱۳۵

نشان داده شده است. مطابق نتایج به‌دست‌آمده، با وجود این که روش TINCD از نتایج ۱۱ روش دیگر تشخیص کمپلکس از روی شبکه PPI و ۵ روش تشخیص کمپلکس‌ها از داده‌های آزمایشگاهی TAP استفاده می‌کند، اما هم‌چنان روش پیشنهادی مقاله در اکثر سنجنده‌های ارزیابی به نتایج بهتری دست یافته است. به طور خاص، یکی از نقاط ضعف روش TINCD، تعداد بسیار زیاد کمپلکس‌های گزارش شده توسط آن است که منجر به کاهش مقدار سنجنده صحّت شده است [۳۲]. نتایج فوق، تأیید دیگری بر ضرورت پویاسازی فرآیند تشخیص و کاهش نوفهٔ شبکه PPI است و توانایی مدل پیشنهادی را نشان می‌دهد که توانسته است با وجود عدم استفاده از داده‌های TAP، به‌دقت بیشتری دست یابد. این نتایج، به ازای شبکه PPI ورودی DIP و مجموعه محک CYC2008 به‌دست‌آمده است. هم‌چنین، یادآوری می‌گردد که نتایج روش TINC، بر مبنای مجموعه کمپلکس‌های گزارش شده توسط آن‌ها به



	صحت	بازخوانی	معیار f	حساسیت	مقدار پیش بینی مثبت	دقت
GA-DCT+SSW+CAMWI	0.78	0.8	0.79	0.7	0.76	0.73
TINCD	0.51	0.73	0.6	0.73	0.72	0.72

شکل ۸-۷: مقایسه مدل پیشنهادی و روش TINCD

ازای همین مجموعه داده‌ها می‌باشد.

۵. نتیجه‌گیری

با توجه به نقش کمپلکس‌های پروتئینی در انجام بسیاری از کارکردهای سلولی موجودات زنده، کشف آن‌ها می‌تواند بـه درک بهتر فرآیندهای سـلولی و توسـعه کاربردهای مبتنی بر مهندسـی زیستی منجر شود. پیشرفت ابزارها و فنّاوری‌های آزمایشگاهی در دهه گذشته منجر به تولید و انباشت داده‌های زیادی از کارکرد اجزاء سلول شده است. این امر، در کنار محدودیت‌های روش‌های آزمایشـگاهی، ضرورت مدل‌سازی سیستمی از طریق بررسی و شناخت سلول در سطح سیستمی و در قالب برهم‌کنش میان اجزاء را اجتناب‌ناپذیر سـاخته اسـت. رویکردهای محاسـباتی ارائه‌شـده تاکنـون به منظـور تشـخیص کمپلکس‌های پروتئینی، به‌طور عمده بر خوشـه‌بندی شبکه PPI تمرکز دارنـد. چالش‌های عمده مسـئله تشـخیص کمپلکس‌های پروتئینی عبارت است از ۱) ضرورت در نظر گرفتن شرایط پویای سلول در فرآیند تشخیص، ۲) مقدار زیاد نوفه از نوع مثبت کاذب در شبکه‌های PPI موجود و ۳) استفاده بیشینه از دانش زیستی موجود در فرآیند تشخیص.

در ایـن مقالـه مدلـی سـه لایـه مبتنی بـر رویکرد زیست‌شناسـی سـامانه‌ها برای تشـخیص کمپلکس‌های پروتئینی از شبکه‌های PPI ارائه‌شده است. طراحی لایه‌های سه‌گانه به‌گونه‌ای است که هر لایه، وظیفه مدیریت و کاهش یکی از چالش‌های سه‌گانه مسئله تشخیص را بر عهده دارد. نوآوری‌های اصلی مدل پیشنهادی عبارت‌اند از:

۱- پیشـنهاد رویکـرد مبتنی بر خوشـه‌بندی دوگانه ماتریس بیان‌ژن، به منظـور در نظر گرفتن ویژگی پویای برهم‌کنش‌هـای بین پروتئینی در تشـخیص کمپلکس‌های پروتئینی.

۲- ارائه روش خوشـه‌بندی دوگانه ماتریس بیان ژن، مبتنی بر الگوریتم ژنتیک بـا تابع برازندگی کارا با قابلیت کشـف الگوهای پیچیده (و سـازگار با مسـئله تشخیص کمپلکس‌هـای پروتئینـی) و مقـاوم به نوفـه موجود در داده‌های بیان ژن.

۳- ارائه روش وزن‌دار کردن شبکه PPI از طریق ترکیب وزن‌دار شباهت معنایی مبتنی بر گراف هستان‌شناسی ژن و شـباهت مبتنی بر پیکربندی شبکه PPI و هم‌چنین، ارائه سنجنده Tsss به منظور محاسبه شباهت معنایی دو پروتئین از روی گراف هستان‌شناسی ژن.

٤- پیـدا کردن هسـته کمپلکس با اسـتفاده از مفاهیم چگالی وزن‌دار و ضریب خوشه‌بندی وزن‌دار، در یک رویه حریصانه، به‌طوری‌که هسـته هر کمپلکـس دارای چگالی وزن‌دار بیشینه می‌باشد.

با توجه به گســتردگی پژوهش‌های انجــام گرفته در لایه‌های سه‌گانه مدل پیشنهادی، محورهای پژوهشی بسیار و متنوعی وجود دارد که توسعه هر یک می‌تواند به افزایش کارایی مدل پیشــنهادی منجر شود. در ادامه به تعدادی از محورهای پژوهشـــی اشاره می شود. برای مثال همچنان، ضرورت ارائه یک روش کارای خوشـــه‌بندی گراف که با ماهیت داده‌های برهم‌کنش سازگاری بیشتری داشته باشد، وجود دارد. برای مثال، ارئه راهکارهای جدید و یا توسعه راهکارهای موجود از طریق مفاهیمی مانند خوشـــه‌بندی فازی گراف و یا خوشه‌بندی مبتنی بر شار عبوری از شبکه می‌تواند مفید باشـــد. هم‌چنین، با توجه به نوفه بسیار زیاد شبکه‌های PPI، اســـتفاده و توسعه روش‌های خوشه‌بندی مقاوم به نوفه یک رهیافت پژوهشی می‌باشد. همچنین، در مورد ارتقای شبکه PPI، هم‌چنان، استخراج میزان شباهت با اســـتفاده از منابع داده‌ای دیگر مانند داده‌های بیان ژن و یا شبکه‌های متابولیت و ... در ترکیب با سایر سنجنده‌های شباهت، می‌تواند به ارتقای کاراتر شبکه PPI منجر شود.

منابع

1. Bader, G.D. and C.W. Hogue, An automated method for finding molecular complexes in large protein interaction networks. BMC bioinformatics, 2003. 4(1): p. 2.
2.Collins, S.R., et al., Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Molecular & Cellular Proteomics, 2007. 6(3): p. 439-450.
3.Chin, C.H., et al., A hub-attachment based method to detect functional modules from confidence-scored protein interactions and expression profiles. BMC Bioinformatics, 2010. 11: p. S25.
4.Gavin, A.C., et al., Proteome survey reveals modularity of the yeast cell machinery. Nature, 2006. 440: p. 631-636.
5.Jain, S. and G.D. Bader, An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. BMC bioinformatics, 2010. 11(1): p. 562.
6.Rasti, S. and C. Vogiatzis, A survey of computational methods in protein–protein interaction networks. Annals of Operations Research, 2019. 276(1-2): p. 35-87.
7.Zahiri, J., et al., Protein complex prediction: A survey. Genomics, 2019.
8.Liu, G., L. Wong, and H.N. Chua, Complex discovery from weighted PPI networks. Bioinformatics, 2009. 25(15): p. 1891-1897.
9.Wang, J., et al., Protein complex detection algorithm based on multiple topological characteristics in PPI networks. Information Sciences, 2019. 489: p. 78-92.
10.Lo, K., et al., Integrating external biological knowledge in the construction of regulatory networks from time-series expression data. BMC systems biology, 2012. 6(1): p. 101.
11.Altaf-Ul-Amin, M., et al., Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC bioinformatics, 2006. 7(1): p. 207.
12.Adamcsek, B., et al., CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics, 2006. 22(8): p. 1021-1023.
13.Jiang, P. and M. Singh, SPICi: a fast clustering algorithm for large biological networks. Bioinformatics, 2010. 26(8): p. 1105-1111.
14.Enright, A.J., S. Van Dongen, and C.A. Ouzounis, An efficient algorithm for large-scale detection of protein families. Nucleic acids research, 2002. 30(7): p. 1575-1584.
15.Nepusz, T., H. Yu, and A. Paccanaro, Detecting overlapping protein complexes in protein-protein interaction networks. Nature methods, 2012. 9(5): p. 471-472.
16.Wu, M., et al., A core-attachment based method to detect protein complexes in PPI networks. BMC bioinformatics, 2009. 10(1): p. 169.
17.Leung, H.C., et al., Predicting protein complexes from PPI data: a core-attachment approach. Journal of Computational Biology, 2009. 16: p. 133-144.
18.King, A.D., N. Pržulj, and I. Jurisica, Protein complex prediction via cost-based clustering. Bioinformatics, 2004. 20(17): p. 3013-3020.
19.Sharan, R., et al., Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. Journal of Computational Biology, 2005. 12(6): p. 835-846.
20.Ou-Yang, L., et al., A two-layer integration framework for protein complex detection. BMC bioinformatics, 2016. 17(1): p. 100.
21.Li, M., et al., Towards the identification of protein complexes and functional modules by integrating PPI network and gene expression data. BMC bioinformatics, 2012. 13(1): p. 109.
22.Ou-Yang, L., et al., Detecting temporal protein complexes from dynamic protein-protein interaction networks. BMC bioinformatics, 2014. 15(1): p. 335.
23.Lakizadeh, A., S. Jalili, and S.-A. Marashi, PCD-GED: Protein complex detection considering PPI Dynamics based on time series gene expression data. Journal of theoretical biology, 2015. 378: p. 31-38.
24.Lakizadeh, A., S. Jalili, and S. Marashi, CAMWI: Detecting protein complexes using weighted clustering coefficient and weighted density. Computational biology and chemistry, 2015. 58: p. 231-240.
25.Lakizadeh, A. and S. Jalili, BiCAMWI: a genetic-based biclustering algorithm for detecting dynamic protein complexes. PloS one, 2016. 11(7).
26.Wang, J., et al., Recent advances in clustering methods for protein interaction networks. BMC genomics, 2010. 11(S3): p. S10.
27.Flores, J.L., et al., A new measure for gene expression biclustering based on non-parametric correlation. Computer methods and programs in biomedicine, 2013. 112(3): p. 367-397.
28.Yu, Y. and Z. Zheng, Protein Complex Identification Based on Weighted PPI Network with Multi-Source Information. Journal of theoretical biology, 2019.
29.Chatr-Aryamontri, A., et al., The BioGRID interaction database: 2013 update. Nucleic acids research, 2012. 41(D1): p. D816-D823.
30.Madeira, S.C. and A.L. Oliveira, A polynomial time biclustering algorithm for finding approximate expression patterns in gene expression time series. Algorithms for Molecular Biology, 2009. 4(1): p. 8.
31.Pu, S., et al., Up-to-date catalogues of yeast protein complexes. Nucleic acids research, 2009. 37(3): p. 825-831.
32.Beer, L.A. and D.W. Speicher, Protein detection in gels using fixation. Current protocols in protein science, 2018. 91(1): p. 10.5. 1-10.5. 20.