مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره ششم شماره اول. ص ۵۸-۴۶، ۱۳۹۸

یک روش ترکیبی اندازه‌های مرکزیت و خواص زیستی برای بهبود تشخیص کمپلکس‌های پروتئینی در شبکه‌های PPI وزنی

عبدالکریم الهی[۱]، سید مرتضی بابامیر[۲*]

• دریافت مقاله: ۱۳۹۷/۶/۱۱ • پذیرش مقاله: ۱۳۹۷/۱۱/۲۷

مقدمه: در شبکه‌های برهمکنش پروتئینی، یک کمپلکس گروهی از پروتئین‌ها است که موجب فرآیند زیستی می‌شوند. شناسایی درست کمپلکس‌ها می‌تواند به فهم بهتر عملکرد سلول‌ها کمک کند تا در اهداف درمانی مانند کشف دارو مورد استفاده قرار گیرد. یکی از روش‌های متداول برای شناسایی کمپلکس‌ها در شبکه‌های برهمکنش پروتئینی، خوشه‌بندی است؛ اما هدف این پژوهش یافتن روشی جدید برای شناسایی دقیق‌تر کمپلکس‌ها است.

روش: در این مطالعه توسعه‌ای–کاربردی از شبکه‌های پروتئینی مخمر و انسان استفاده شد. مجموعه‌های داده‌ای مخمر به نام‌های MIPS،DIP و Krogan به ترتیب دارای ۴۹۳۰ گره و ۱۷۲۰۱ برهمکنش، ۴۵۶۴ گره و ۱۵۱۷۵ برهمکنش و ۲۶۷۵ گره و ۷۰۸۴ برهمکنش و مجموعه داده‌ای انسان دارای ۳۷۴۳۷ برهمکنش است. الگوریتم پیشنهادی و الگوریتم‌های مشهور در شناسایی کمپلکس‌های پروتئینی بر روی مجموعه‌های داده‌ای اجرا شده‌اند و کمپلکس‌های پیش‌بینی شده با مجموعه داده‌های معیار CYC2008 و CORUM مورد مقایسه قرار گرفتند.

نتایج: در این تحقیق روش جدیدی از دسته روش‌های مبتنی بر هسته و پروتئین‌های الحاقی جهت تشخیص کمپلکس‌های پروتئینی استفاده شد که دارای کارایی بالایی در تشخیص بود. هرچه قدر تشخیص کمپلکس‌ها دقیق‌تر باشد، می‌توان پروتئین‌های دخیل در یک فرآیند زیستی را درست‌تر تشخیص داد. معیارهای ارزیابی نشان داد که روش پیشنهادی، بهبود قابل‌توجهی نسبت به دیگر روش‌ها دارد.

نتیجه‌گیری: با توجه به نتایج به دست آمده مشاهده شد که روش پیشنهادی تعداد مناسبی از کمپلکس‌های پروتئینی را شناسایی نمود و بیشترین نسبت معنی‌داری زیستی را در همکاری عملکردی پروتئین‌ها دارد.

کلید واژه‌ها: کمپلکس‌های پروتئینی، شبکه برهمکنش پروتئینی، اندازه‌های مرکزیت، پروتئین‌های اساسی، الگوریتم مبتنی بر هسته و پروتئین‌های جانبی

• **ارجاع:** الهی عبدالکریم، بابامیر سید مرتضی. یک روش ترکیبی اندازه‌های مرکزیت و خواص زیستی برای بهبود تشخیص کمپلکس‌های پروتئینی در شبکه‌های PPI وزنی. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۸؛ ۶(۱): ۵۸-۴۶.

۱. دانشجوی دکتری مهندسی کامپیوتر، گروه مهندسی کامپیوتر، دانشکده برق و کامپیوتر، دانشگاه کاشان، کاشان، ایران
۲. دکتری مهندسی کامپیوتر، دانشیار، گروه مهندسی کامپیوتر، دانشکده برق و کامپیوتر، دانشگاه کاشان، کاشان، ایران
* **نویسنده مسئول:** کاشان، قطب راوندی، دانشگاه کاشان
• **شماره تماس:** ۰۹۱۳۱۶۳۵۲۱۱
• **Email:** babamir@kashanu.ac.ir

مقدمه

پروتئین‌ها مسئول بسیاری از فعالیت‌ها و رفتارها در سلول هستند. تقریباً ۸۰ درصد پروتئین‌ها در کنار یکدیگر فعالیت می‌کنند و با برهمکنش آن‌ها فرآیند زیستی شکل می‌گیرد. با پیشرفت فناورهای آزمایشگاهی با خروجی انبوه، شبکه‌های برهمکنش پروتئینی بزرگی ایجاد شده‌اند. شبکه برهمکنش پروتئین – پروتئین(Protein-Protein interaction) PPI، شبکه‌ای از برهمکنش‌های فیزیکی در میان پروتئین‌ها است [۱] که پروتئین‌ها به‌عنوان رأس‌ها و برهمکنش‌های فیزیکی بین جفت پروتئین‌ها به‌عنوان یال‌های گراف در نظر گرفته می‌شوند. شبکه‌های PPI، نقش بسیار مهمی در فرآیندهای زیستی شامل کنترل چرخه، جداسازی، تاخوردگی پروتئین، انتقال سیگنال، رونویسی و ترجمه بازی می‌کنند [۲]. شناسایی کمپلکس‌های پروتئینی در شبکه PPI، کمک زیادی به درک مکانیسم فعالیت زیستی، واحدهای اساسی عملکرد و معماری شبکه برهمکنش پروتئینی دارد. کمپلکس‌های پروتئینی، اجتماعی از پروتئین‌ها هستند که در یک زمان و مکان خاص با هم برهمکنش دارند تا یک فرآیند زیستی خاصی را انجام دهند. تاکنون الگوریتم‌های محاسباتی مختلفی با دیدگاه‌های متفاوت برای عمل خوشه‌بندی روی شبکه‌های برهمکنش پروتئین-پروتئین مطرح شده‌اند. معمولاً در این روش‌ها، شبکه‌های برهمکنش پروتئین-پروتئین با استفاده از گراف مدل شده و زیر گراف چگال به عنوان کمپلکس پروتئینی در نظر گرفته می‌شوند. بر همین اساس اکثر این روش‌ها از مفاهیم نظریه گراف استفاده می‌کنند و به‌دلیل نادیده گرفتن خواص زیستی دقت بالایی ندارند. در سال‌های اخیر محققان تلاش کرده‌اند تا با دخالت دادن برخی اطلاعات زیستی، الگوریتم‌های بهتری ارائه دهند [۳،۴]. روش MCL (Markov CLuster)، شبکه PPI را بر اساس شبیه‌سازی گام‌های تصادفی خوشه‌بندی می‌کند [۵]. بعضی روش‌ها مانند (Molecular Complex Detection) MCODE زیرگراف‌های چگال را به عنوان کمپلکس‌های پروتئینی پیش‌بینی می‌کند [۶] . الگوریتم‌های (a software tool for network cluster) Cfinder [۷] و CMC (clustering-based on maximal cliques) [۸] با یافتن و ادغام کلیک‌ها (زیر گراف کامل که هر دو رأس آن به وسیله یک یال به هم متصل شده باشند [۹]) به شناسایی کمپلکس‌های پروتئینی می‌پردازند.

بر اساس مطالعاتی که بر روی کمپلکس‌های Yeast انجام شد، کمپلکس پروتئینی از دو بخش اصلی هسته و پروتئین‌های الحاقی به هسته تشکیل شده‌اند [۱۰]. هسته نقش اصلی کمپلکس را بر عهده دارد و پروتئین‌های الحاقی نقش کمک‌کننده در پروتئین‌های هسته را دارند. در یک کمپلکس پروتئینی بین پروتئین‌های هسته نسبت به سایر پروتئین‌ها، برهمکنش بیشتری وجود دارد و هسته دارای چگالی بیشتری نسبت به قسمت‌های دیگر کمپلکس دارد. با توجه به دقت بالای روش‌های این دسته نسبت به روش‌های دیگر، روش پیشنهادی این پژوهش نیز مبتنی بر روش‌های هسته و پروتئین‌های الحاقی است. در این زمینه، Srihari و همکاران روشی به نام MCL-Caw معرفی کردند که پروتئین‌ها را در خوشه‌های تولید شده از الگوریتم MCL به پروتئین‌های هسته و الحاقی بر اساس ارتباطات آن‌ها گروه‌بندی می‌کنند و سپس تنها این پروتئین‌های دسته‌بندی شده را انتخاب می‌کند درحالی‌که پروتئین‌های باقی‌مانده (نویزی) را حذف می‌کند [۱۱]. Peng و همکاران روشی به نام (weighted PageRank-Nibble algorithm and core-attachment structure) WPNCA ارائه کردند که شبکه PPI وزنی را به چندین زیر گراف متصل به وسیله الگوریتم PageRank-Nibble تقسیم کرده و سپس کمپلکس‌های پروتئینی در هر زیر گراف بر اساس ساختار core-attachment به دست می‌آیند [۱۲]. از مشهورترین روش‌های دیگر این دسته می‌توان به CORE [۱۳] COACH(core-attachment based method)، [۱۴] PEWCC(predicts protein complexes based on the concept of weighted clustering coefficient) و [۱۵] اشاره کرد. روش COACH در دو گام کار می‌کند. در گام اول ناحیه‌های بسیار متصل از شبکه به نام هسته‌های اولیه کشف می‌شوند سپس این مناطق به وسیله همسایه‌های بسیار متصل گسترش می‌یابند [۱۴]. ازجمله مشکلات این روش، انتخاب تصادفی رأس مرکزی برای ساخت زیر گراف هسته، انتخاب نادقیق پروتئین‌های الحاقی، عدم وجود وزن در برهمکنش میان پروتئین‌ها و داشتن نویزها در شبکه برهمکنش پروتئینی است که در روش پیشنهادی مطالعه حاضر این مشکلات حل شده است. روش PEWCC نیز از دو مرحله اصلی، تعیین اطمینان برهمکنش‌ها، حذف برهمکنش‌های نویزی و کشف کمپلکس‌ها با استفاده از ضریب خوشه‌بندی تشکیل شده است [۱۵]؛ اما در روش پیشنهادی با استفاده از ترکیب اندازه مرکزیت با خواص زیستی و فیلتر رأس‌های نویزی، رأس‌های مرکزی هسته به صورت دقیق‌تری

انتخاب می‌شوند. با استفاده از چگالی وزنی و انتخاب بهتر پروتئین‌های الحاقی، شناسایی کمپلکس‌های پروتئینی دقیق‌تر می‌شود. در این پژوهش، داده‌های ورودی با استفاده از آنتولوژی ژنی، وزن‌گذاری می‌شوند، اطمینان برهمکنش بین پروتئین‌ها اندازه‌گیری شده، نویزها حذف می‌گردند سپس برای یافتن پروتئین‌های اساسی به عنوان Seed در مرکز هسته‌های اولیه، پروتئین‌های نویزی فیلتر می‌شوند. بر اساس مراکز ابتدایی و چگالی وزنی، هسته کمپلکس‌ها و پروتئین‌های الحاقی تشکیل می‌شوند. البته دو گام ابتدایی از رویکرد پیشنهادی یعنی وزن-گذاری شبکه ورودی و اندازه اهمیت پروتئین‌ها در شبکه‌های وزنی از پژوهش قبلی نویسندگان این مقاله استفاده شد. در پژوهش قبلی برای افزایش دقت تشخیص پروتئین‌های اساسی، کلاس‌بند چندگانه که شامل اندازه‌های مرکزیت برای خاصیت‌های محلی و سراسری شبکه و ویژگی‌های زیستی است، به کار گرفته شد [۱۶]. الگوریتم پیشنهادی و الگوریتم‌های مشهور در شناسایی کمپلکس-های پروتئینی مانند، (core-attachment based method) COACH [۱۴] ، CORE [۱۳] ، (Clustering algorithm with Overlapping Neighborhood Expansion) ClusterOne [۱۷] ، (Molecular Complex Detection) MCODE [۶] و (Identification of Protein Complexes Algorithm) IPCA [۱۸] بر روی مجموعه‌های داده‌ای(The Database of Interacting Proteins) DIP [۱۹] ، (Munich Information Center for Protein Sequences) MIPS [۲۰] و Krogan [۲۱] اجرا شده‌اند و کمپلکس‌های پیش‌بینی شده با مجموعه داده‌ای استاندارد CYC2008 [۲۲] مورد مقایسه قرار گرفته‌اند. معیارهای ارزیابی نشان می‌دهند که الگوریتم پیشنهادی با کارایی بالاتری کمپلکس‌های پروتئینی را پیش‌بینی می‌کند.

روش

این مطالعه از نوع توسعه‌ای–کاربردی است. برای ارزیابی روش CEBCOA (Composite Biological Features and CEntrality Measures) از ۳ شبکه برهمکنش پروتئین‌های مخمر به نام‌های (Database of Interacting Proteins) DIP [۱۹]، (Munich Information Center for Protein Sequences) MIPS [۲۰] و Krogan [۲۱] استفاده شد. مجموعه داده‌ای DIP دارای ۴۹۳۰ پروتئین و ۱۷۲۰۱ برهمکنش، شبکه Krogan دارای ۲۶۷۵ پروتئین و ۷۰۸۴ برهمکنش و شبکه MIPS دارای ۴۵۶۴ پروتئین و ۱۵۱۷۵ برهمکنش می‌باشد. در این ارزیابی از مجموعه‌های کمپلکس‌های واقعی به نام CYC 2008 استفاده شد [۲۲] . این مجموعه شامل ۴۲۸ کمپلکس مخمر می‌باشد. برای ساخت شبکه‌های وزن‌دار، مقدار مشابهت معنایی بین جفت پروتئین‌ها بر اساس آنتولوژی ژنی محاسبه شد. شبکه‌های غیر جهت‌دار و وزن‌دار با گراف G=(V,E,W) نشان داده شد که $V = \{v_1, v_2, ..., v_n\}$ مجموعه‌ای از رأس‌ها و $E = \{e_1, e_2, ..., e_n\}$ مجموعه‌ای از یال‌ها و W وزن هر یال می‌باشد. وزن $W_{i,j}$، قدرت برهمکنش بین رأس $V i$ و $V j$ را نشان می‌دهد. این محاسبات با استفاده از GFD-Net [۲۳] پلاگینی از ابزار Cytoscape در وزن‌گذاری مجموعه‌های داده‌ای به کار گرفته شد و برهمکنش‌های مثبت کاذب در این شبکه‌ها حذف شدند. برنامه کاربردی GFD-Net از Cytoscape برای بصری‌سازی و آنالیز عدم مشابهت عملکردی در شبکه‌های پروتئینی (ژنی) طراحی شده است و می‌تواند یک شبکه پروتئینی را بر اساس آنتولوژی ژنی (Gene Ontology) GO و کمیتی از تفاوت عملکردشان وزن‌گذاری نماید.

در این پژوهش الگوریتم جدیدی به نام CEBCOA معرفی شد. این روش دارای گام‌های اساسی زیر می‌باشد.

۱- محاسبه شباهت معنایی برهمکنش جفت پروتئین‌ها بر اساس آنتولوژی ژنی
۲- حذف برهمکنش‌ها با شباهت معنایی کم
۳- فیلتر نویزهای پروتئینی مانند Active site [۲۴] ، bridge [۲۵] .
۴- کشف پروتئین‌های اساسی به عنوان Seed برای مراکز ابتدایی کمپلکس‌ها
۵- پیدا کردن هسته‌های اولیه
۶- حذف هسته‌های تکراری
۷- اضافه کردن پروتئین‌های الحاقی به هر هسته برای تشکیل یک کمپلکس پیش‌بینی شده
۸- حذف کمپلکس‌های تکراری در مرحله ۷ و به دست آوردن کمپلکس‌های نهایی

در حوزه تئوری گراف و تحلیل شبکه، تعدادی از اندازه‌ها برای تعیین اهمیت رأس‌ها وجود دارند که به تعریف چند اندازه مرکزیت مشهور در زیر پرداخته شد.

تعریف ۱: مرکزیت درجه وزنی

مرکزیت درجه وزنی ($DCW(i)$) از رأس i، مجموع وزن لبه-هایی است که رأس i را به همسایگان آن متصل می‌کند (رابطه ۱) [۲۶].

$$DCW(i) = \sum_{j \in N_i} W_{i,j} \qquad (۱)$$

که N_i مجموعه همسایگان رأس i است.

تعریف ۲: مرکزیت بینابینی وزنی

مرکزیت بینابینی وزنی ($BCW(i)$) رأس i برابر است با متوسط مقدار کوتاه‌ترین مسیرهایی که از طریق رأس i می‌گذرند (معادله ۲) [۲۶].

$$BCW(i) = \sum_s \sum_t \frac{\sigma_{st}(i)}{\sigma_{st}}, s \neq t \neq i \qquad (۲)$$

به‌طوری‌که σ_{st} تعداد کل کوتاه‌ترین مسیرها بین رأس‌های s و t است و $\sigma_{st}(i)$ تعداد کوتاه‌ترین مسیرهایی است که از رأس s به رأس t از طریق رأس i می‌گذرند.

تعریف ۳: مرکزیت اتصال متوسط محلی وزنی

مرکزیت اتصال متوسط محلی وزنی (LACW(i)) رأس i در یک گراف وزنی G به صورت رابطه ۳ است [۲۷].

$$LACW(i) = \frac{\sum_{s \in N_u} \sum_{t \in N_s \cap N_u} W(s,t)}{|N_u|} \qquad (۳)$$

به‌طوری‌که N_u مجموعه‌ای از تمام همسایگان رأس u، $|N_u|$ تعداد همسایگان آن و W(s,t) وزن لبه اتصالی رأس s به رأس t که رأس s همسایه رأس t است.

معمولاً برای مشخص نمودن اهمیت هر رأس از خاصیت‌های توپولوژی گراف استفاده شد. یکی از مشهورترین اندازه‌ها برای تشخیص (رأس‌های با اهمیت) پروتئین‌های اساسی، درجه رأس است. مرکزیت درجه برای اندازه محلی شبکه استفاده می‌شود؛ اما این معیار ارتباطات شبکه‌ای را به صورت ضعیف نشان می‌دهد و در اکثر پژوهش‌ها اشاره شده است که اندازه مرکزیت به تنهایی نمی‌تواند تمامی پروتئین‌های اساسی به ویژه پروتئین‌های اساسی با اتصال–کم را به درستی پیش‌بینی کند؛ بنابراین در این مطالعه با استفاده از ترکیب قوانین زیستی و معیارهای مرکزیت، دقت تشخیص پروتئین های اساسی با استفاده از معادله ۴ افزایش می‌یابد [۱۶] .

$$COBCEM(x) = a \times DWNN(x) + b \times MCLUS(x) + c \times Ess(x) + d \times LACW(x) \qquad (۴)$$

که a، b، c و d ضرایبی برای نشان دادن اهمیت هر کدام از قسمت‌های معادله را نشان می‌دهد. برای به دست آوردن مقادیر مناسب a، b، c و d آزمایش‌هایی توسط کلاس‌بند رگرسیون لجستیک دودویی انجام شد. تابع DWNN برای نشان دادن تأثیر رأس‌های همسایه در فاصله دو سطحی، تابع MCLUS برای اندازه‌گیری خاصیت بیولوژی تجمع پروتئین‌های اساسی در یک ماژول و شناسایی بهتر پروتئین‌های اساسی با اتصال–کم، تابع Ess برای اندازه‌گیری خاصیت بیولوژی ارتباط پروتئین‌های اساسی با پروتئین‌های مهم دیگر و شناسایی بهتر پروتئین‌های اساسی با اتصال–کم و Local Average Connectivity Weighted) LACW(method نیز معیار مرکزیتی است که با استفاده از آن ترکیبی از معیارهای مرکزیت محلی و سراسری شبکه در تشخیص بهتر پروتئین‌های اساسی استفاده شده است [۱۶] . وقتی دو رأس دارای قدرت و تعداد برهمکنش یکسانی باشند؛ اما ویژگی‌های توپولوژی مختلفی داشته باشند آنگاه تعداد و قدرت برهمکنش‌ها نمی‌توانند مقایسه خوبی در شناسایی پروتئین‌های اساسی داشته باشند. به همین دلیل در این مطالعه همسایگی سطح‌های بیشتر برای تشخیص تفاوت توپولوژی رأس‌ها مورد بررسی قرار داده شد [۱۶] .(معادلات ۵ و ۶)

$$DW(x) = \sum_{y \in adj(x)} N(y) \qquad (۵)$$

$$DWNN(y) = \sum_{x \in adj(y)} W_{y,x} \times DW(y) \qquad (۶)$$

که adj(x)، مجموعه همسایه‌های نزدیک به رأس x و N(y) مجموع مرکزیت درجه وزنی ضربدر مرکزیت بینابینی رأس y و همسایه‌های غیر مستقیم آن در دو گام بعدی را نشان می‌دهد. ضریب $W_{y,x}$ نیز تأثیر قدرت برهمکنش در اهمیت پروتئین اساسی را نشان می‌دهد که از طریق وزن لبه‌ها مقداردهی می‌شود. در شبکه PPI تعدادی ماژول عملکردی وجود دارد که نقش کلیدی در عملکرد بیولوژی بازی می‌کنند و پروتئین‌های اساسی تمایل دارند در این ماژول‌های عملکردی قرار گیرند. با به‌کارگیری افزونه MCODE در نرم‌افزار Cytoscape عملیات خوشه‌بندی روی مجموعه داده‌ای انجام شد سپس رتبه هر پروتئین نسبت به خوشه‌های موجود محاسبه

گردید. اگر CP={ $cp_1, cp_2, \ldots, cp_n$ } مجموعه‌ای از کمپلکس‌های پروتئینی و پروتئین x در کمپلکس‌هایی وجود داشته باشد آنگاه مجموع درجه وزنی آن پروتئین نسبت به پروتئین‌های داخل آن کمپلکس‌ها به عنوان اندازه پروتئین x (MCLUS (x)) به دست می‌آید (معادله ۷).

$$MCLUS(x) = \sum W_{x,y} \mid x, y \in \{cp_1, cp_2, \ldots, cp_n\} \quad and \quad y \in neighbor(x) \qquad (۷)$$

از دیگر ویژگی‌های زیستی پروتئین‌های اساسی، می‌توان به ارتباط آن‌ها با پروتئین‌های مهم دیگر اشاره نمود. برای اندازه‌گیری این ویژگی زیستی به صورت Ess (x) نشان داده شد (معادله ۸).

$$Ess(x) = \sum W_{x,y} \mid y \in neighbor(x) \; and \quad y \in \{\text{Important Proteins}\} \qquad (۸)$$

که Important Proteins از الگوریتم ۱ محاسبه شد.

Algorithm 1. Important protein algorithm
Input: G (V, E, W)
Output: the set of important proteins (SIP)

1- Initialize the dataset of the important protein SIP= {∅}.
2- Calculate the weighted degree of each protein in the dataset.
3- Find the protein with the largest weighted degree (p-max) in the dataset and add it to the SIP.
4- Remove p-max and its neighboring proteins from the dataset.
5- Return to step three and repeat until the dataset is empty.
6- Output the SIP.

با استفاده از روش رگرسیون لجستیک ترکیب اندازه‌های مرکزیت با سه معیار، همسایگی در سطح‌های دوم و سوم (DWNN)، MCLUS و Ess آزمایش شد و معیار مرکزیت اتصال متوسط محلی وزن‌دار (LACW) با سطح معنی‌داری خوبی نسبت به اندازه‌های مرکزیت دیگر نتیجه شده است [۱۶] . سپس با رتبه‌بندی پروتئین‌ها، قضاوتی در مورد اساسی بودن آن‌ها انجام شد. قبل از رتبه‌بندی، تمامی مقادیر اندازه‌ها به وسیله روش نرمال‌سازی ماکزیمم–مینیمم نرمال شده و به ارزش‌هایی بین ۰,۰ و ۱,۰ تبدیل خطی شدند. اگر $\max(X_{ij})$ ، $\min(X_{ij})$ به ترتیب ماکزیمم و مینیمم مقدار ویژگی jام باشد و X_{ij} مقدار اندازه پروتئین i در مرکزیت و ویژگی j باشد آنگاه مقدار نرمال مقادیر اندازه‌ها طبق رابطه ۹ به دست می‌آید:

$$NORMAL(X_{i,j}) = \frac{X_{ij} - \min(X_{ij})}{\max(X_{ij}) - \min(X_{ij})} \qquad (۹)$$

در ادامه مراحل روش پیشنهادی، به حذف پروتئین‌های نویزی مانند Active Site [۲۴] و Bridge [۲۵] پرداخته شد. با وجود چنین پروتئین‌هایی تشخیص پروتئین‌های اساسی از دقت خوبی برخوردار نخواهند بود و معمولاً با استفاده از اندازه‌های مرکزیت آن‌ها به اشتباه به عنوان پروتئین اساسی در نظر گرفته می‌شوند. درصورتی‌که نقش خاصی در کمپلکس‌های پروتئینی ندارند و دقت تشخیص پروتئین‌های اساسی را کاهش می‌دهند. پروتئین Active Site به صورت تصادفی در شبکه‌های برهمکنش پروتئینی پخش شده‌اند و خوشه‌های جدا از هم را به یکدیگر متصل می‌کنند. این پروتئین‌ها ارتباطات ضعیفی با پروتئین‌های هاب دارند [۲۴] . پروتئین bridge، پروتئینی در شبکه PPI است که زیر گراف همسایه‌ها به وسیله آن از هم جدا شده‌اند [۲۵] . شناسایی این نوع پروتئین‌ها در تصفیه و فیلتر شبکه PPI کمک می‌کند. در نهایت با حذف پروتئین‌های نویزی، پروتئین‌هایی که مقدار COBCEM (Composite (Biological Features and Centrality Measures آن‌ها بزرگ‌تر از α باشد به عنوان مراکز ابتدایی (Seed) در نظر گرفته می‌شود (الگوریتم ۲). به وسیله تنظیم α می‌توان تعداد کمپلکس پیش‌بینی شده را کنترل نمود. مقادیر زیاد برای α کمپلکس‌های پروتئینی کمتر را موجب می‌شود. مقدار α از متوسط ۵۰ درصد بالایی مقادیر COBCEM برای انتخاب مراکز ابتدایی انجام شد.

Algorithm 2. Seed generation algorithm
Input: G (V, E, W), DCW, BCW, LACW
Output: the set of candidate proteins for core centers

1- Seeds ← an empty set.
2- For each protein ∈ V(G) do
3- Compute its DWNN, MCLUS and Ess function

4- Compute its COBCEM function as Eq.4
5- End For
6- For each protein P ∈ V(G) do
7- If COBCEM(P)> α then Seeds ← Seeds ⋃ P
8- End For

در ادامه برای کشف هسته کمپلکس، شبکه وزن‌دار G=(V,E,W) به عنوان ورودی در نظر گرفته و مجموعه C به صورت زیر تشکیل شد.

$$C = \{(Gc_V, S_V) \mid V \in Seed \quad and \quad S_V = \phi\} \qquad (۱۰)$$

در معادله ۱۰، Gc_V گراف هسته رأس v است که از روابط زیر به دست می‌آید.

$$V^{GC}V = V_W(G_V) \qquad (۱۱)$$

که V_W مجموعه رئوس هسته در یک گراف وزن‌دار (معادله ۱۲)، G_V گراف همسایگی رأس $v \in Seed$ و S_V مجموعه‌ای از رئوس که ابتدا تهی است.

$$V_W(G) = \{u \in v \mid W(u) > W_m(G)\} \qquad (۱۲)$$

در معادله ۱۲، $W_m(G)$ میانگین وزنی یال‌های شبکه است. با استفاده از الگوریتم ۳، مجموعه هسته‌های اولیه در شبکه ورودی تشکیل شد.

Algorithm 3: predict preliminary Cores
Input: The set C, d threshold
Output: The set of preliminary cores P

1- P = ∅
2- For each $(Gc_V, S_V) \in C$
3- If ($D_w(Gc_V) \geq d$)
4- Each vertex of Gc_V is added to S_V and Gc_V is appended to C
5- Else
6- Sub-graph Gc_V is decomposed by removing the vertices of $V_w(Gc_V)$ from V^{Gc_V}.
7- For each connected component of decomposition and $V_w(G_C)$ are considered as a pair add to C.

که D_w چگالی وزن‌دار (معادله ۱۳) و d حد آستانه برای کنترل چگالی هسته‌های کشف شده است [۲۸].

$$Dw(G) = \frac{2 \times \sum(u,v) \in E^{W(u,v)}}{|V|(|V|-1)} \qquad (۱۳)$$

پس از کشف هسته‌های اولیه و حذف هسته‌های تکراری با توجه به میزان همپوشانی آن‌ها (الگوریتم ۴) مجموعه هسته‌های نهایی F به دست می‌آید.

Algorithm 4: Redundancy Filtering
Input: the set of Preliminary Cores P, t threshold
Output: The set of final cores F

1- F= ϕ
2- For each core graph $G_C \in$ P
3- G_C'=argmax NA(G_C', G_C) | $G_C \in$F
4- If NA(G_C', G_C)<t
5- insert G_C into F
6- ELSE
7- If den(G_C) * $|V^{G_C}|$ ≥ den (G_C') * $|V^{G_C'}|$
8- replace G_C' with G_C in F

که NA(G,G') میزان همپوشانی دو گراف G, G' از رابطه میل به همسایگی (معادله ۱۴) محاسبه می‌شود [۲۹].

$$NA(G,G') = \frac{|V \cap V'|^2}{|V| \times |V'|} \qquad (۱۴)$$

و t، حد آستانه برای کنترل میزان همپوشانی بین هسته‌های کشف شده است. در ادامه برای یافتن نزدیکی بین پروتئین‌های الحاقی V و هسته‌ها از رابطه ۱۵ استفاده می‌شود.

$$A(c) = \{V \in N(G_C) \mid \sigma(V, G_C) \succ \varepsilon\} \qquad (۱۵)$$

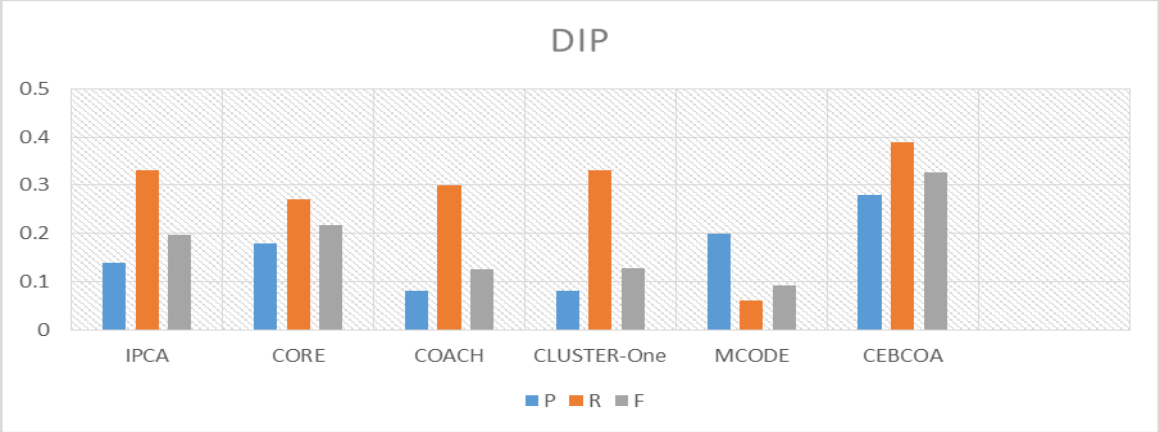
که $N(G_c)$ مجموعه همسایه‌های هسته، ε حد آستانه میزان نزدیکی بین پروتئین‌های الحاقی و هسته و $\sigma(V, G_C)$ اندازه نزدیکی بین رأس v و زیر گراف هسته G_C می‌باشد (معادله ۱۶).

$$\sigma(V, G_C) = \Sigma_{u \in V^{G_C}} \frac{W(u,v)}{|V^{G_C}|} \qquad (۱۶)$$

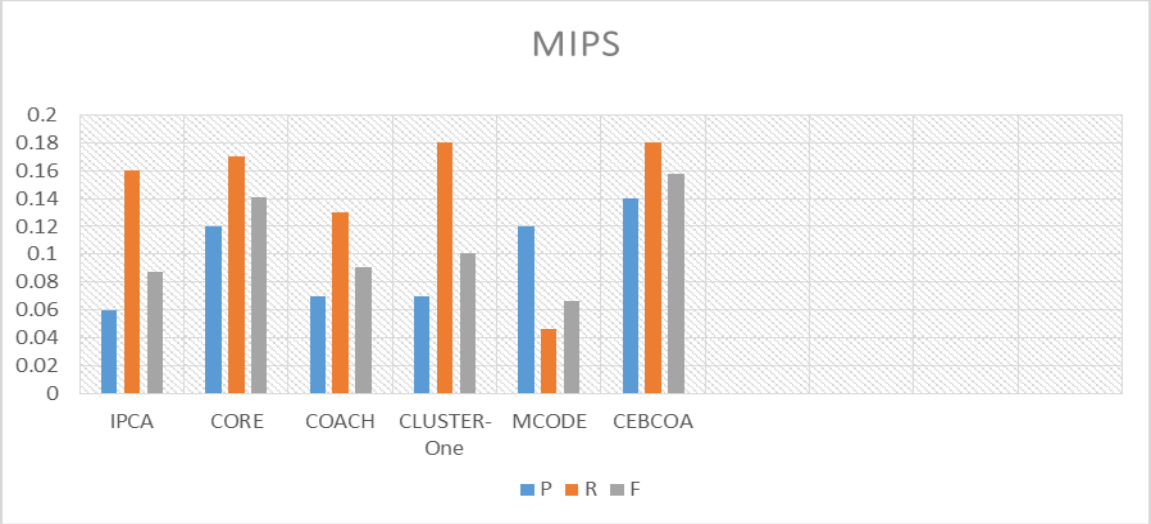
در نهایت پس از یافتن پروتئین‌های الحاقی و ترکیب آن‌ها با هسته‌ها، کمپلکس‌ها تشکیل می‌شوند.

نتایج

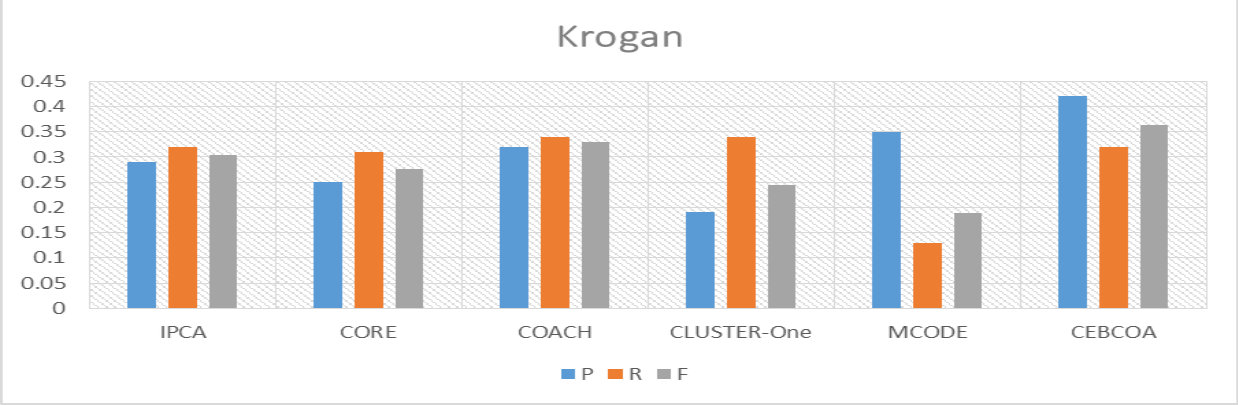
شکل‌های ۱، ۲ و ۳ نتیجه اجرای روش پیشنهادی با روش‌های مشهور، روی سه شبکه DIP، Krogan، MIPS را نشان می‌دهد. همان‌طور که در شکل‌ها مشاهده می‌شود، روش CEBCOA عملکرد بهتری از تمامی روش‌های مشهور مطرح شده دارد.



شکل ۱: معیارهای Precision، Recall و F-measure الگوریتم‌های مورد مقایسه بر روی مجموعه داده‌ای DIP



شکل ۲: معیارهای Precision، Recall و F-measure الگوریتم‌های مورد مقایسه بر روی مجموعه داده‌ای MIPS



شکل ۳: معیارهای Precision، Recall و F-measure الگوریتم‌های مورد مقایسه بر روی مجموعه داده‌ای Krogan

برای مقایسه CEBCOA با روش‌های دیگر از معیارهای Precision، Recall، F–measure و P_value طبق معادلات ذیل استفاده شد [۳۰-۳۲].

$$precision = \frac{\left|\left\{K \middle| K \in \kappa, \exists R \in \Re, \alpha(R,K) \geq \theta\right\}\right|}{|\kappa|} \quad (۱۷)$$

$$\text{Re}\,call = \frac{\left|\left\{R \middle| R \in \Re, \exists K \in \kappa, \alpha(R,K) \geq \theta\right\}\right|}{|\Re|} \quad (۱۸)$$

$$F-measure = \frac{2 \times precision \times \text{Re}\,call}{precision + \text{Re}\,call} \quad (۱۹)$$

$$P\_value = 1 - \sum_{i=0}^{k-1} \frac{\binom{m}{i}\binom{N-m}{|V^p|-i}}{\binom{N}{|V^p|}} \quad (۲۰)$$

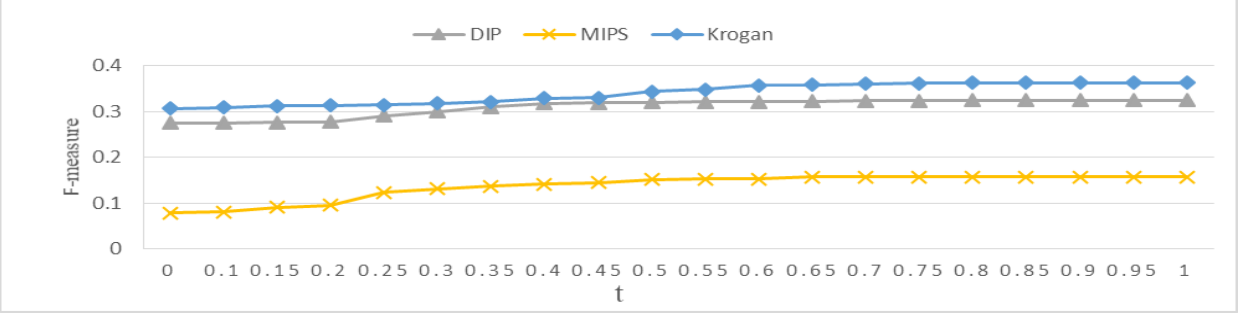
مجموعه $\kappa = \{k_1, k_2, \ldots, k_n\}$: مجموعه کمپلکس‌های کشف شده است.

مجموعه $\Re = \{R_1, R_2, \ldots, R_m\}$: مجموعه کمپلکس‌های واقعی است.

$\alpha(R,K)$: میزان همپوشانی بین دو کمپلکس است.

θ: حد آستانه همپوشانی بین کمپلکس‌های واقعی و کمپلکس‌های کشف شده است. به‌طورمعمول مقدار θ برابر ۰/۵ در نظر گرفته می‌شود [۳۱].

معیار P_value: برای ارزیابی معنی‌داری بیولوژیکی کمپلکس‌های پیش‌بینی شده استفاده می‌شود. اگر k پروتئین در کمپلکس پیش‌بینی شده $P = (V^p, E^p)$ از نظر عملکرد یکسان فرض شوند و در یک گروه عملکردی با m عضو قرار گیرند آنگاه مقدار P_value از رابطه ۲۰ به دست می‌آید. در این رابطه، N تعداد کل پروتئین‌ها در شبکه PPI است. برای نمونه جدول ۱ مقایسه نتایج معنی‌داری بیولوژیکی در مجموعه داده‌ای DIP را نشان می‌دهد. با توجه به نتایج جدول، کمپلکس‌های پیش‌بینی شده در روش پیشنهادی ۸۸/۴۸٪ معنی‌دار هستند و روش CEBCOA بیشترین نسبت را در میان روش‌های مطرح شده به دست آورده است. هر چقدر مقدار P_value کمتر باشد ارزش آماری بالاتری دارد و نشان می‌دهد که پروتئین‌های کمپلکس به صورت تصادفی کنار هم قرار نگرفتند [۳۲]. اگر مقدار P_value از کمپلکس‌های پیش‌بینی شده کمتر از ۰/۰۱ باشد آنگاه معنی‌دار خواهند بود [۳۲]. با استفاده از ابزار GoTermFinder [۳۳] می‌توان مقدار P_value را از ساختار آنتولوژی ژنی به دست آورد. در مقایسه الگوریتم پیشنهادی با دیگر الگوریتم‌ها، الگوریتمی که تعداد کمپلکس‌های معنی‌دار بیشتری را پیش‌بینی نماید، بهتر است. حد آستانه مقدار میل به همسایگی t در زیر الگوریتم Redundancy – filtering برای کنترل میزان همپوشانی بین هسته‌های کشف شده استفاده می‌شود. همان‌طور که در شکل ۴ نشان داده شد با افزایش مقدار t مقدار F-measure نیز افزایش می‌یابد. وقتی که t بزرگ‌تر از ۰/۸ باشد، F-measure در مجموعه داده‌ای DIP ثابت باقی می‌ماند. در مجموعه داده‌ای MIPS وقتی t بزرگ‌تر از ۰/۶۵ شود F-measure با مقادیر ثابتی باقی می‌ماند. در Krogan هم هنگامی‌که t بزرگ‌تر از ۰/۸ باشد F-measure تمایل به ثابت ماندن دارد؛ بنابراین وقتی مقدار حد آستانه میل به همسایگی به ۰/۸ تنظیم شود، تمامی مجموعه‌های داده‌ای دارای بالاترین مقدار F-measure و ثابت باقی می‌مانند.



شکل ٤: تأثیر مقادیر مختلف حد آستانه میل به همسایگی t و مقادیر F-measure در مجموعه‌های داده‌ای

جدول ۱: درصد معنی‌داری آماری کمپلکس‌های پیش بینی شده از الگوریتم‌های مشهور در مجموعه داده‌ای DIP

الگوریتم	تعداد کمپلکس‌های پیش‌بینی شده	تعداد کمپلکس‌های معنی‌دار	نسبت
IPCA	۸۰۶	۵۵۹	۶۹/۳۵
CORE	۳۴۹	۲۰۱	۵۷/۵۹
COACH	۷۳۰	۵۸۴	۸۰
ClusterOne	۱۰۳۵	۷۱۳	۶۸/۸۸
MCODE	۶۵	۵۰	۷۶/۹۲
CEBCOA	۴۹۵	۴۳۸	۸۸/۴۸

در این مطالعه d (حد آستانه چگالی) برای کشف هسته‌های اولیه به کار می‌رود و چگالی وزنی مربوط به هر گراف هسته با استفاده از رابطه ۲۱ به‌دست می‌آید.

$$Dave(Gc_V) = \frac{2 \times |E^{Gc_V}| \times W_{ave}}{|V^{Gc_V}|(|V^{Gc_V}| - 1)} \qquad (۲۱)$$

همچنین ε (حد آستانه مقدار نزدیکی)، معیاری برای کشف پروتئین‌های الحاقی با استفاده از وزن برهمکنش‌ها و محاسبه نزدیکی بین یک پروتئین الحاقی و یک هسته تعریف می‌شود. بالاترین مقدار F– measure زمانی به دست آمده که مقدار ε برابر ۰/۴۹۷ بوده است. بر همین اساس به‌طور پیش فرض ε برابر ۰/۵ در نظر گرفته شد. در این آزمایش مقدار t برابر ۰/۸ بود. علاوه بر آزمایش‌های فوق امروزه محققان از مجموعه داده‌های Homo Sapiens برای ارزیابی بهتر روش‌هایشان استفاده می‌کنند. در این پژوهش نیز از مجموعه داده‌ای PPI Human [۳۴] با ۳۷۴۳۷ برهمکنش و مجموعه کمپلکس‌های معیار CORUM [۳۵] به عنوان استاندارد طلایی شامل ۱۸۴۳ کمپلکس انسانی استفاده شد.

بحث و نتیجه‌گیری

در این مطالعه، ارزیابی جامعی از الگوریتم‌های مشهور مطرح و روش پیشنهادی در زمینه پیش‌بینی کمپلکس‌های پروتئینی انجام گرفت. نتایج آزمایشگاهی نقاط قوت و ضعف هر الگوریتم را نشان داد. روش پیشنهادی CEBCOA برای تشخیص کمپلکس‌های پروتئینی، ابتدا با شناسایی دقیق‌تر از مراکز ابتدایی، هسته‌ها را تشکیل می‌دهد و سپس با محاسبه چگالی وزنی هر کدام از پروتئین‌های الحاقی با هسته‌ها، به تشکیل کمپلکس‌ها ادامه می‌دهد. کار با داده‌های PPI محدودیت‌های بسیاری مانند نویزهای زیاد داده‌ای، تعداد زیاد کمپلکس‌های کوچک همراه با تعدادی کمپلکس بزرگ، پروتئین‌های دارای عملکرد همپوشان در چندین کمپلکس را دارد [۳۶]. در این مطالعه با ساخت شبکه وزنی، حذف برهمکنش‌های مثبت کاذب و فیلتر پروتئین‌های ActiveSite و Bridge نتایج بهتری جهت پیش‌بینی کمپلکس‌ها صورت گرفت. پس از مقایسه‌های مختلف CEBCOA با دیگر الگوریتم‌های مشهور نتایج نشان داد که روش پیشنهادی با بسیاری از الگوریتم‌های مشهور به لحاظ مقادیر F-measure، Recall، Precision و P_value عملکرد بالاتری داشته است. به‌طورکلی با توجه به ایده‌ها در روش پیشنهادی در مجموعه‌های داده‌ای DIP، MIPS و Krogan بهترین بود و در مجموعه داده‌ای انسان نیز عملکرد تقریباً خوبی داشت. الگوریتم CEBCOA روی شبکه DIP نتایج بسیار خوبی دارد و مقدار Precision، Recall و F-measure آن از همه الگوریتم‌های دیگر بالاتر است. مقدار F-measure روش CEBCOA برابر ۰/۳۲۶ است درصورتی‌که برای (Identification of Protein Complexes Algorithm) IPCA [۱۸]، CORE [۱۳]، COACH(Core-Attachment based method) [۱۴]، (Clustering algorithm with Overlapping Neighborhood Expansion) ClusteOne [۱۷] و MCODE(Molecular Complex Detection) [۶] به ترتیب برابر ۰/۱۹۶، ۰/۲۱۶، ۰/۱۲۶، ۰/۱۲۸ و ۰/۰۹۲ است. در شبکه MIPS الگوریتم CEBCOA در هر سه معیار بررسی شده نسبت به سایر الگوریتم‌ها کارایی بهتری دارد. در شبکه MIPS روش CEBCOA مقدار Precision، Recall و F–measure به ترتیب برابر ۰/۱۴، ۰/۱۸، ۰/۱۵۷ است درصورتی‌که مقدار F-measure دیگر روش‌ها IPCA [۱۸]، CORE [۱۳]، COACH [۱۴]، ClusterOne [۱۷] و MCODE [۶] به ترتیب برابر ۰/۰۸۷، ۰/۱۴، ۰/۰۹۱، ۰/۱۰۱ و ۰/۰۶۷ است. برای شبکه Krogan، مقدار Precision و F–measure الگوریتم CEBCOA از

سایر الگوریتم‌ها بیشتر بوده که این مقادیر برای CEBCOA به ترتیب برابر با ۰/۴۲ و ۰/۳۶۳ است، درحالی‌که برای روش-های IPCA [۱۸] ، CORE [۱۳] ، COACH [۱۴] ، ClusterOne [۱۷] و MCODE [۶] مقادیر Precision و F-measure به ترتیب برابر با(۰/۳۰۴، ۰/۲۹)، (۰/۲۷۷، ۰/۲۵)، (۰/۳۳۰، ۰/۳۲)، (۰/۱۹، ۰/۲۴۴) و (۰/۳۵، ۰/۱۸۹) است؛ اما مقدار Recall برای CEBCOA، IPCA [۱۸] ، CORE [۱۳] ، COACH [۱۴] ، ClusterOne [۱۷] و MCODE [۶] به ترتیب برابر ۰/۳۲، ۰/۳۲، ۰/۳۱، ۰/۳۴، ۰/۳۴ و ۰/۱۳ است که حدود ۰/۰۲ کمتر از الگوریتم‌های Cluster-One و COACH و برابر با الگوریتم IPCA می‌باشد. با توجه به انتخاب دقیق‌تر مراکز هسته‌ها به عنوان Seed از طریق پروتئین‌های اساسی، انتخاب پروتئین‌هایی با ارتباط عملکردی بهتر در هسته‌ها، به‌کارگیری چگالی وزنی قابل‌قبول در اتصال پروتئین‌های الحاقی به هسته‌ها و استفاده از آنتولوژی ژنی، باعث شده که روش پیشنهادی قادر به فیلتر کمپلکس‌های کم اطمینان و شناسایی کمپلکس‌های پروتئینی با معنی‌داری زیستی بالا شود؛ بنابراین هسته‌هایی که به وسیله CEBCOA شناسایی می‌شوند می‌بایستی بعضی عملکردهای مشترک را به اشتراک بگذارند و این مطابقت بیشتری با تعریف هسته کمپلکس‌ها دارد و همین مسئله باعث پیدا کردن کمپلکس‌های دقیق‌تر می‌شود درصورتی‌که در روش‌های دیگر عدم استفاده از مراکز هسته‌ای دقیق و کشف هسته‌های اولیه نادقیق، پیش‌بینی کمپلکس‌های بی‌معنی زیستی را بیشتر کرده است و باعث شده که آن‌ها دارای F-measure پایین‌تری نسبت به روش پیشنهادی باشند. در شبکه انسان، CEBCOA مقدار ۰/۲۲۷، ۰/۱۵۵، ۰/۱۸۴ برای Precision، Recall و F-measure به دست آورد. درصورتی‌که مقدار F-measure روش‌های IPCA [۱۸] ، CORE [۱۳] ، COACH [۱۴] ، ClusterOne [۱۷] و MCODE [۶] به ترتیب برابر ۰/۷۷، ۰/۱۸۳، ۰/۱۹۵، ۰/۱۶۲ و ۰/۰۷ به دست آمده است. با توجه به مقادیر اشاره شده CEBCOA بهترین عملکرد را بعد از روش COACH دارد. در شبکه انسان نیز روش این مطالعه تقریباً مقدار خوبی از F-measure را دارد که نسبت به COACH، ۰/۰۱۱ کمتر و بهتر از بقیه روش‌های مطرح شده است. درصد معنی‌داری کمپلکس-های پیش‌بینی شده در روش‌های مختلف می‌تواند به عنوان مقدار کارآمدی آن روش مورد استفاده قرار گیرد. با توجه به نتایج جدول ۱، کمپلکس‌های پیش‌بینی شده در روش پیشنهادی ۸۸/۴۸٪ معنی‌دار هستند و این روش بیشترین نسبت را در میان روش‌ها به دست آورده است. این نسبت به ترتیب به اندازه ۱۹٪، ۳۰٪ ، ۸٪، ۱۹ ٪ و ۱۱ ٪ بالاتر از روش‌های IPCA [۱۸] ، CORE [۱۳] ، COACH [۱۴] ، ClusterOne [۱۷] و MCODE [۶] می‌باشد؛ بنابراین روش پیشنهادی قادر است عملکرد مناسبی در تشخیص کمپلکس‌های پروتئینی داشته باشد به‌طوری‌که معنی بیولوژیکی بهتری نسبت به دیگران حفظ کرده است. درصورتی‌که روش‌هایی مانند Core و ClusterOne به علت پیش‌بینی کمپلکس‌های کوچک با مقادیر بزرگ P_value، نتایج ضعیفی دارند. کمپلکس‌های پروتئینی با اندازه بزرگ با احتمال بالایی دارای P_value کوچک‌تری هستند. الگوریتم‌های بسیاری برای تشخیص کمپلکس‌های پروتئینی پیشنهاد شده‌اند؛ اما هنوز در پیش‌بینی کمپلکس‌های پروتئینی از مجموعه داده‌های مختلف، دقیق و کارآمد نیستند. همچنین موفقیت‌های هر کدام از روش‌ها تا حد زیادی بستگی به تکنیک‌های آزمایشگاهی زیست‌شناسان دارد که مجموعه داده‌ای بیولوژیکی قابل‌اطمینان را برای محاسبات فراهم نمایند؛ بنابراین وقتی دانشمندان کامپیوتر و زیست‌شناسان با هم کار می‌کنند برای دانشمندان کامپیوتر با اضافه شدن داده‌هایی که زیست‌شناسان فراهم می‌کنند، راه‌هایی مطمئن و کافی برای کاویدن دانش جدید از PPI فراهم می‌شود. در آینده می‌توان با استفاده از توابع مشابهت همسایگی، برهمکنش‌های منفی کاذب را پیش‌بینی کرده تا با حذف آن‌ها شبکه‌هایی کامل‌تر ایجاد کرد. آنگاه با اطلاعات اضافی مانند اطلاعات دامنه‌ای و داده‌های بیان‌ژنی کمپلکس‌های پروتئینی را با دقت بیشتری پیش‌بینی کرده و با تحقیق معنی‌داری بیولوژیکی کمپلکس‌ها، درک کامل‌تری از ویژگی‌های آن‌ها کشف شود.

با توجه به نقش کمپلکس‌های پروتئینی در فرآیندهای زیستی، شناسایی کمپلکس‌های پروتئینی یک مسئله مهم در محاسبات زیستی است. در مطالعه حاضر روشی جدید مبتنی بر هسته و پروتئین‌های الحاقی برای تشخیص کمپلکس‌های پروتئینی در شبکه‌های وزن‌دار ارائه شد. اندازه شباهت معنایی بین جفت پروتئین‌ها برای تخمین اطمینان برهمکنش‌های پروتئینی بر اساس آنتولوژی ژنی استفاده شد. با استفاده از اندازه‌های ترکیبی مرکزیت و ویژگی‌های زیستی، اهمیت هر کدام از پروتئین‌ها به دست آمد. آنگاه با مرتب‌سازی نزولی اندازه‌ها، پروتئین‌های مرکزی Seed تشخیص داده شد و سپس با فیلتر پروتئین‌های نویزی از پروتئین‌های اساسی، دقت انتخاب پروتئین‌های مرکزی در مراکز هسته بیشتر شد. در ادامه هسته

اولیه کمپلکس‌ها با چگالی وزنی بالا مشخص و سپس پروتئین‌های الحاقی به هسته اضافه شده‌اند. نتایج آزمایشگاهی نشان داد که دقت تشخیص کمپلکس‌های پروتئینی در روش پیشنهادی نسبت به دیگر روش‌ها از بهبود قابل‌توجهی برخوردار است. همچنین کمپلکس‌های پروتئینی پیش‌بینی شده در روش پیشنهادی از درصد معنی‌داری بیولوژیکی بیشتری برخوردار هستند.

تعارض منافع

این مقاله دارای تعارض منافع نمی‌باشد.

References

1. Srihari S, Leong HW. A survey of computational methods for protein complex prediction from protein interaction networks. J Bioinform Comput Biol 2013;11(2):1230002.

2. Tu S, Chen R, Xu L. A binary matrix factorization algorithm for protein complex prediction. Proteome Sci 2011;9 Suppl 1:S18.

3. Zhang XF, Dai DQ, Ou-Yang L, Yan H. Detecting overlapping protein complexes based on a generative model with functional and topological properties. BMC Bioinformatics 201415:186

4. Ou-Yang L, Zhang XF, Dai DQ, Wu MY, Zhu Y, Liu Z, Yan H. Protein complex detection based on partially shared multi-view clustering. BMC Bioinformatics 201617:371

5. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 2002;30(7):1575-84.

6. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 2003;4:2.

7. Adamcsek B, Palla G, Farkas IJ, Derenyi I, Vicsek T. CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics 2006;22(8):1021-3.

8. Liu G, Wong L, Chua HN. Complex discovery from weighted PPI networks. Bioinformatics 2009;25(15):1891-7.

9. Junker BH, Schreiber F. Analysis of Biological Networks. 1th ed. New York: Wiley-Interscience; 2011.

10. Dezso Z, Oltvai ZN, Barabasi AL. Bioinformatics analysis of experimentally determined protein complexes in the yeast Saccharomyces cerevisiae. Genome Res 2003;13(11):2450-4.

11. Srihari S, Ning K, Leong HW. MCL-CAw: a refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure. BMC Bioinformatics 2010;11:504.

12. Peng W, Wang J, Zhao B, Wang L. Identification of protein complexes using weighted PageRank-Nibble algorithm and core-attachment structure. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2015; 12(1):179-92.

13. Srihari S, Yong CH, Patil A, Wong L. Methods for protein complex prediction and their contributions towards understanding the organisation, function and dynamics of complexes. FEBS Lett 2015 14;589(19 Pt A):2590-602.

14. Price T, Pena FI, 3rd, Cho YR. Survey: Enhancing protein complex prediction in PPI networks with GO similarity weighting. Interdiscip Sci 2013;5(3):196-210.

15. Zaki N, Efimov D, Berengueres J. Protein complex detection using interaction reliability assessment and weighted clustering coefficient. BMC Bioinformatics 2013;14:163.

16. Elahi A, Babamir SM. Identification of essential proteins based on a new combination of topological and biological features in weighted protein-protein interaction networks. IET Syst Biol 2018;12(6):247-57.

17. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. Nature Methods 2012; 9(5): 471.

18. Li M, Chen JE, Wang JX, Hu B, Chen G. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. BMC Bioinformatics 2008;9:398.

19. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, et al. IntAct--open source resource for molecular interaction data. Nucleic Acids Res 2007;35(Database issue):D561-5.

20. Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, et al. MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res 2004;32(Database issue):D41-4.

21. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 2006;440(7084):637-43.

22. Pu S, Wong J, Turner B, Cho E, Wodak SJ. Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res 2009;37(3):825-31.

23. Diaz-Diaz N, Diaz-Montana JJ. GFD-Net: a novel apporach for analyzing the functional dissimilarity of gene networks. In6th Argentinian Conference on Bioinformatics and Computational Biology 2015 Oct 14.

24. Csermely P. Creative elements: network-based predictions of active centres in proteins and cellular and social networks. Trends Biochem Sci 2008;33(12):569-76.

25. Zaki N, Berengueres J, Efimov D. Detection of protein complexes using a protein ranking algorithm. Proteins: Structure, Function, and Bioinformatics. 2012; 80(10):2459-68.

26. Li M, Wang J, Wang H, Pan Y. Essential proteins discovery from weighted protein interaction networks. International Symposium on Bioinformatics Research

and Applications; 2010 May 23; Berlin: Springer; 2010. p. 89-100.
27. Tang Y, Li M, Wang J, Pan Y, Wu FX. CytoNCA: a cytoscape plugin for centrality analysis and evaluation of protein interaction networks. Biosystems 2015;127:67-72.
28. Yu Y, Wang X, Lin L, Sun C, Wang X. Detecting protein complexes based on sequence information in the weighted protein–protein interaction network. Journal of Computational and Theoretical Nanoscience. 2012;9(10):1565-70.
29. Wu M, Li X, Kwoh CK, Ng SK. A core-attachment based method to detect protein complexes in PPI networks. BMC Bioinformatics 2009;10:169.
30. Ahn J, Lee DH, Yoon Y, Yeu Y, Park S. Improved method for protein complex detection using bottleneck proteins. BMC Med Inform Decis Mak 2013;13 Suppl 1:S5.
31. Habibi M, Eslahchi C, Wong L. Protein complex prediction based on k-connected subgraphs in protein interaction network. BMC Syst Biol 2010;4:129.
32. Ma X, Gao L. Predicting protein complexes in protein interaction networks using a core-attachment algorithm based on graph communicability. Information Sciences 2012;189:233-54.
33. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, et al. GO:TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 2004;20(18):3710-5.
34. Keretsu S, Sarmah R. Weighted edge based clustering to identify protein complexes in protein-protein interaction networks incorporating gene expression profile. Comput Biol Chem 2016;65:69-79.
35. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, et al. CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res 2010;38(Database issue):D497-501.
36. Liu Q, Song J, Li J. Using contrast patterns between true complexes and random subgraphs in PPI networks to predict unknown protein complexes. Sci Rep 2016;6:21223.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2019; 6(1):46-58

A Combination Method of Centrality Measures and Biological Properties to Improve Detection of Protein Complexes in Weighted PPI Networks

Elahi Abdolkarim[1], Babamir Seyed Morteza[2]*

• **Received:** 2 Sep, 2018 • **Accepted:** 16 Feb, 2019

Introduction: In protein-protein interaction networks (PPINs), a complex is a group of proteins that allows a biological process to take place. The correct identification of complexes can help better understanding of the function of cells used for therapeutic purposes, such as drug discoveries. One of the common methods for identifying complexes in the PPINs is clustering, but this study aimed to identify a new method for more accurate identification of complexes.

Method: In this study, Yeast and Human PPINs were investigated. The Yeast datasets, called DIP, MIPS, and Krogan, contain 4930 nodes and 17201 interactions, 4564 nodes and 15175 interactions, and 2675 nodes and 7084 interactions, respectively. The Human dataset contains 37437 interactions. The proposed and well-known methods have been implemented on datasets to identify protein complexes. Predicted complexes were compared with the CYC2008 and CORUM benchmark datasets. The evaluation criteria showed that the proposed method predicts PPINs with higher efficiency.

Results: In this study, a new method of the core-attachment methods was used to detect protein complexes enjoying high efficiency in the detection. The more precise the detection method is, the more correct we can identify the proteins involved in biological process. According to the evaluation criteria, the proposed method showed a significant improvement in the detection method compared to the other methods.

Conclusion: According to the results, the proposed method can identify a sufficient number of protein complexes, among the highest biological significance in functional cooperation with proteins.

Keywords: Protein complexes, Protein interaction network, Centrality measures, Essential protein, Core-attachment Algorithm

• **Citation:** Elahi A, Babamir SM. A Combination Method of Centrality Measures and Biological Properties to Improve Detection of Protein Complexes in Weighted PPI Networks. Journal of Health and Biomedical Informatics 2019; 6(1): 46-58. [In Persian]

1. Ph.D. Student in Computer, Faculty of Electrical and Computer Engineering, University of Kashan, Kashan, Iran
2. Ph.D. in Computer, Associate Professor, Faculty of Electrical and Computer Engineering, University of Kashan, Kashan, Iran.

***Correspondence**: University of Kashan, Ghotb-e-Ravandi Blvd, Kashan, Iran
• **Tel:** 09131635211 • **Email:** Babamir@kashanu.ac.ir