

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

عبدالکریم الهی^۱، سید مرتضی بابامیر^{۲*}

• پذیرش مقاله: ۱۳۹۷/۱۱/۲۷

• دریافت مقاله: ۱۳۹۷/۶/۱۱

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 قابل توجهی 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

• **Email:** babamir@kashanu.ac.ir

• **شماره تماس:** ۰۹۱۳۱۶۳۵۲۱۱

مقدمه

پروتئین ها مسئول بسیاری از فعالیت ها و رفتارها در سلول هستند. تقریباً ۸۰ درصد پروتئین ها در کنار یکدیگر فعالیت می کنند و با برهمکنش آن ها فرآیند زیستی شکل می گیرد. با پیشرفت فناوری های آزمایشگاهی با خروجی انبوه، شبکه های برهمکنش پروتئینی بزرگی ایجاد شده اند. شبکه برهمکنش پروتئین - پروتئین (PPI (Protein-Protein interaction) شبکه ای از برهمکنش های فیزیکی در میان پروتئین ها است [۱] که پروتئین ها به عنوان رأس ها و برهمکنش های فیزیکی بین جفت پروتئین ها به عنوان یال های گراف در نظر گرفته می شوند. شبکه های PPI، نقش بسیار مهمی در فرآیندهای زیستی شامل کنترل چرخه، جداسازی، تاخوردگی پروتئین، انتقال سیگنال، رونویسی و ترجمه بازی می کنند [۲]. شناسایی کمپلکس های پروتئینی در شبکه PPI، کمک زیادی به درک مکانیسم فعالیت زیستی، واحدهای اساسی عملکرد و معماری شبکه برهمکنش پروتئینی دارد. کمپلکس های پروتئینی، اجتماعی از پروتئین ها هستند که در یک زمان و مکان خاص با هم برهمکنش دارند تا یک فرآیند زیستی خاصی را انجام دهند. تاکنون الگوریتم های محاسباتی مختلفی با دیدگاه های متفاوت برای عمل خوشه بندی روی شبکه های برهمکنش پروتئین - پروتئین مطرح شده اند. معمولاً در این روش ها، شبکه های برهمکنش پروتئین - پروتئین با استفاده از گراف مدل شده و زیر گراف چگال به عنوان کمپلکس پروتئینی در نظر گرفته می شوند. بر همین اساس اکثر این روش ها از مفاهیم نظریه گراف استفاده می کنند و به دلیل نادیده گرفتن خواص زیستی دقت بالایی ندارند. در سال های اخیر محققان تلاش کرده اند تا با دخالت دادن برخی اطلاعات زیستی، الگوریتم های بهتری ارائه دهند [۳،۴]. روش MCL (Markov CLuster)، شبکه PPI را بر اساس شبیه سازی گام های تصادفی خوشه بندی می کند [۵]. بعضی روش ها مانند (Molecular Complex) MCODE (Detection) زیرگراف های چگال را به عنوان کمپلکس های پروتئینی پیش بینی می کند [۶]. الگوریتم های (software tool for network cluster) Cfinder [۷] و (clustering-based on maximal cliques) CMC [۸] با یافتن و ادغام کلیک ها (زیر گراف کامل که هر دو رأس آن به وسیله یک یال به هم متصل شده باشند [۹]) به شناسایی کمپلکس های پروتئینی می پردازند. بر اساس مطالعاتی که بر روی کمپلکس های Yeast انجام شد، کمپلکس پروتئینی از دو بخش اصلی هسته و پروتئین های

الحاقی به هسته تشکیل شده اند [۱۰]. هسته نقش اصلی کمپلکس را بر عهده دارد و پروتئین های الحاقی نقش کمک کننده در پروتئین های هسته را دارند. در یک کمپلکس پروتئینی بین پروتئین های هسته نسبت به سایر پروتئین ها، برهمکنش بیشتری وجود دارد و هسته دارای چگالی بیشتری نسبت به قسمت های دیگر کمپلکس دارد. با توجه به دقت بالای روش های این دسته نسبت به روش های دیگر، روش پیشنهادی این پژوهش نیز مبتنی بر روش های هسته و پروتئین های الحاقی است. در این زمینه، Srihari و همکاران روشی به نام MCL-Caw معرفی کردند که پروتئین ها را در خوشه های تولید شده از الگوریتم MCL به پروتئین های هسته و الحاقی بر اساس ارتباطات آن ها گروه بندی می کنند و سپس تنها این پروتئین های دسته بندی شده را انتخاب می کند درحالی که پروتئین های باقی مانده (نویزی) را حذف می کند [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 کمپلکس های پروتئینی در هر زیر گراف بر اساس ساختار core-attachment به دست می آیند [۱۲]. از مشهورترین روش های دیگر این دسته می توان به CORE [۱۳]، COACH (core-attachment based method) [۱۴] و PEWCC (predicts protein complexes based on the concept of weighted clustering coefficient) [۱۵] اشاره کرد. روش COACH در دو گام کار می کند. در گام اول ناحیه های بسیار متصل از شبکه به نام هسته های اولیه کشف می شوند سپس این مناطق به وسیله همسایه های بسیار متصل گسترش می یابند [۱۴]. از جمله مشکلات این روش، انتخاب تصادفی رأس مرکزی برای ساخت زیر گراف هسته، انتخاب نادقیق پروتئین های الحاقی، عدم وجود وزن در برهمکنش میان پروتئین ها 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 برهمکنش ها، حذف برهمکنش های نویزی و کشف کمپلکس ها با استفاده از ضریب خوشه بندی تشکیل شده است [۱۵]؛ اما در روش پیشنهادی با استفاده از ترکیب اندازه مرکزیت با خواص زیستی و فیلتر رأس های نویزی، رأس های مرکزی هسته به صورت دقیق تری

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 نویری 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 کمپلکس‌های پروتئینی مانند، (core-attachment based method) Clustering [۱۳]، CORE [۱۴]، COACH algorithm with Overlapping Neighborhood Molecular [۱۷]، ClusterOne (Expansion) [۱۷]، MCODE (Complex Detection) [۶] (Identification of Protein Complexes Algorithm) و IPCA [۱۸] بر روی مجموعه‌های داده‌ای (The Database of Interacting Proteins) DIP [۱۹]، 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 (Composite Biological Features and Centrality) CEBCOA (Measures of Interacting Proteins) (Database of Interacting Proteins) DIP [۱۹]، 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 $V=\{v_1,v_2,\dots,v_n\}$ مجموعه‌ای از رأس‌ها و $E=\{e_1,e_2,\dots,e_n\}$ 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.

$$COBCEM(x) = a \times DWNN(x) + b \times MCLUS(x) + c \times Ess(x) + d \times LACW(x) \quad (4)$$

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)

$$DW(x) = \sum_{y \in adj(x)} N(y) \quad (5)$$

$$DWNN(y) = \sum_{x \in adj(y)} W_{y,x} \times DW(y) \quad (6)$$

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 مقداری 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

تعریف ۱: مرکزیت درجه وزنی

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

$$DCW(i) = \sum_{j \in N_i} W_{i,j} \quad (1)$$

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].

$$BCW(i) = \sum_s \sum_t \frac{\sigma_{st}(i)}{\sigma_{st}}, s \neq t \neq i \quad (2)$$

به‌طوری که 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].

$$LACW(i) = \frac{\sum_{s \in N_u} \sum_{t \in N_s \cap N_u} W(s,t)}{|N_u|} \quad (3)$$

به‌طوری که 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 پروتئین‌های اساسی با استفاده از معادله ۴ افزایش می‌یابد [۱۶].

آن‌ها انجام شد. قبل از رتبه‌بندی، تمامی مقادیر اندازه‌ها به وسیله روش نرمال‌سازی ماکزیم-مینیم نرمال شده و به ارزش‌هایی بین ۰ و ۱، تبدیل خطی شدند. اگر $\min(X_{ij})$ و $\max(X_{ij})$ به ترتیب ماکزیم و مینیم مقدار ویژگی i ام 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:

$$NORMAL(X_{ij}) = \frac{X_{ij} - \min(X_{ij})}{\max(X_{ij}) - \min(X_{ij})} \quad (9)$$

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. در صورتی که نقش خاصی در کمپلکس‌های پروتئینی ندارند و دقت تشخیص پروتئین‌های اساسی را کاهش می‌دهند. پروتئین Active Site به صورت تصادفی در شبکه‌های بهمکنش 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 $\leftarrow$ an empty set.
- 2- For each protein $\in V(G)$ do
- 3- Compute its DWNN, MCLUS and Ess function

گردید. اگر $CP = \{cp_1, cp_2, \dots, cp_n\}$ 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).

$$MCLUS(x) = \sum W_{x,y} | x, y \in \{cp_1, cp_2, \dots, cp_n\} \text{ and } y \in neighbor(x) \quad (7)$$

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).

$$Ess(x) = \sum W_{x,y} | y \in neighbor(x) \text{ and } y \in \{ \text{Important Proteins} \} \quad (8)$$

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 $SIP = \{\emptyset\}$.
- 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

که D_w چگالی وزن دار (معادله ۱۳) و d حد آستانه برای کنترل چگالی هسته های کشف شده است [۲۸].

$$D_w(G) = \frac{2 \times \sum_{(u,v) \in E^{W(u,v)}}}{|V|(|V| - 1)} \quad (13)$$

پس از کشف هسته های اولیه و حذف هسته های تکراری با توجه به میزان همپوشانی آنها (الگوریتم ۴) مجموعه هسته های نهایی F به دست می آید.

Algorithm 4: Redundancy Filtering

Input: the set of Preliminary Cores P , t threshold

Output: The set of final cores F

- 1- $F = \emptyset$
- 2- For each core graph $G_C \in P$
- 3- $G_C' = \text{argmax } NA(G_C', G_C) \mid G_C \in F$
- 4- If $NA(G_C', G_C) < t$
- 5- insert G_C into F
- 6- ELSE
- 7- If $\text{den}(G_C) * |V^{G_C}| \geq \text{den}(G_C') * |V^{G_C'}|$
- 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 می شود [۲۹].

$$NA(G, G') = \frac{|V \cap V'|^2}{|V| \times |V'|} \quad (14)$$

و t ، حد آستانه برای کنترل میزان همپوشانی بین هسته های کشف شده است. در ادامه برای یافتن نزدیکی بین پروتئین های الحاقی V و هسته ها از رابطه ۱۵ استفاده می شود.

$$A(c) = \{V \in N(G_C) \mid \sigma(V, G_C) > \varepsilon\} \quad (15)$$

که $N(G_C)$ مجموعه همسایه های هسته، ε حد آستانه میزان نزدیکی بین پروتئین های الحاقی و هسته و $\sigma(V, G_C)$ اندازه نزدیکی بین رأس v و زیر گراف هسته G_C می باشد (معادله ۱۶).

$$\sigma(V, G_C) = \sum_{u \in V^{G_C}} \frac{W(u, v)}{|V^{G_C}|} \quad (16)$$

در نهایت پس از یافتن پروتئین های الحاقی و ترکیب آنها با هسته ها، کمپلکس ها تشکیل می شوند.

4- Compute its COBCEM function as Eq.4

5- End For

6- For each protein $P \in V(G)$ do

7- If $\text{COBCEM}(P) > \alpha$ then $\text{Seeds} \leftarrow \text{Seeds} \cup P$

8- End For

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

$$C = \{(G_{C_V}, S_V) \mid V \in \text{Seed and } S_V = \emptyset\} \quad (10)$$

در معادله ۱۰، G_{C_V} گراف هسته رأس v است که از روابط زیر به دست می آید.

$$V^{G_{C_V}} = V_w(G_V) \quad (11)$$

که V_w مجموعه رئوس هسته در یک گراف وزن دار (معادله ۱۲)، G_V گراف همسایگی رأس $v \in \text{Seed}$ و S_V مجموعه ای از رئوس که ابتدا تهی است.

$$V_w(G) = \{u \in v \mid W(u) > W_m(G)\} \quad (12)$$

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

Algorithm 3: predict preliminary Cores

Input: The set C , d threshold

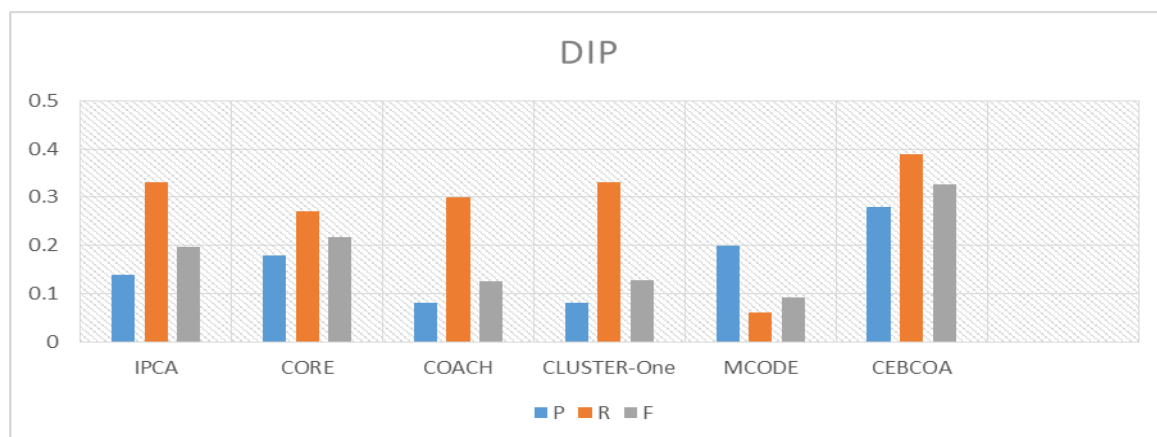
Output: The set of preliminary cores P

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

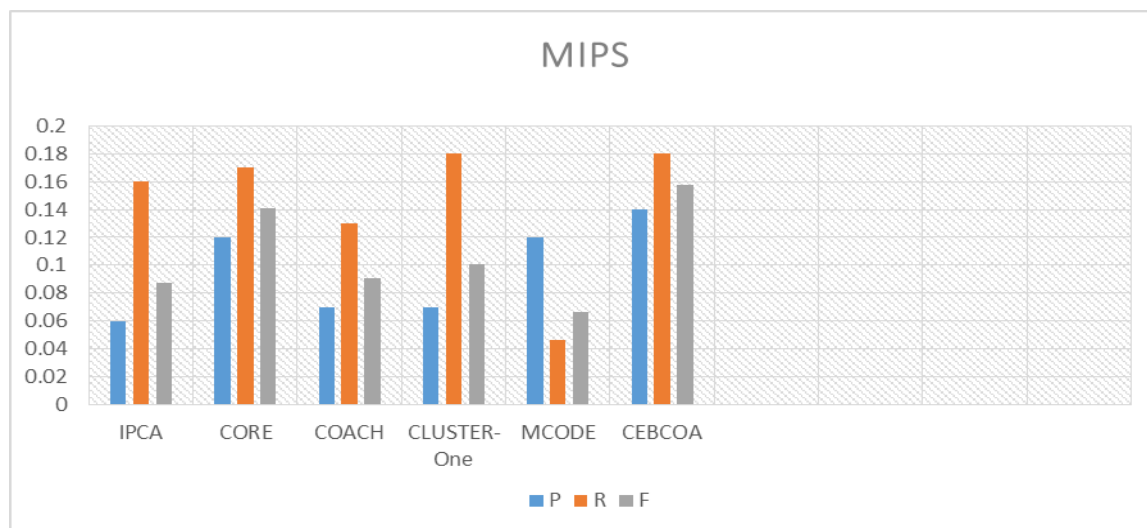
نتایج

شکل‌های ۱، ۲ و ۳ نتیجه اجرای روش پیشنهادی با روش‌های مشهور، روی سه شبکه DIP، Krogan، MIPS را نشان

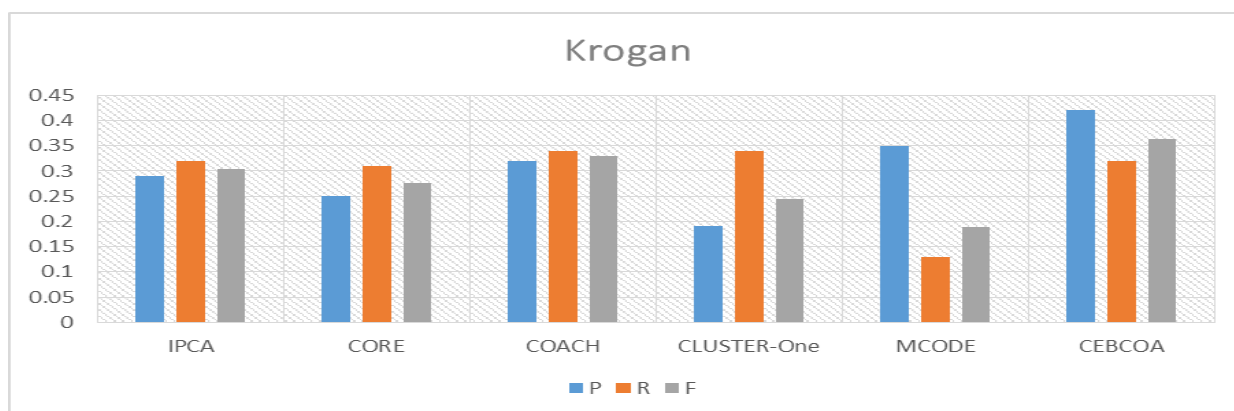
می‌دهد. همان‌طور که در شکل‌ها مشاهده می‌شود، روش CEBCOA عملکرد بهتری از تمامی روش‌های مشهور مطرح شده دارد.



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

کمپلکس های پیش بینی شده استفاده می شود. اگر k پروتئین در کمپلکس پیش بینی شده $P = (VP, EP)$ 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 می آید. در این رابطه، N تعداد کل پروتئین ها در شبکه PPI است. برای نمونه جدول ۱ مقایسه نتایج معنی داری بیولوژیکی در مجموعه داده های DIP را نشان می دهد. با توجه به نتایج جدول، کمپلکس های پیش بینی شده در روش پیشنهادی ۸۸/۴۸٪ معنی دار هستند و روش CEBCOA بیشترین نسبت را در میان روش های مطرح شده به دست آورده است. هر چقدر مقدار P_value کمتر باشد ارزش آماری بالاتری دارد و نشان می دهد که پروتئین های کمپلکس به صورت تصادفی کنار هم قرار نگرفتند [۳۲]. اگر مقدار P_value از کمپلکس های پیش بینی شده کمتر از ۰/۰۱ باشد آنگاه معنی دار خواهند بود [۳۲]. با استفاده از ابزار GoTermFinder [۳۳] می توان مقدار P_value را از ساختار آنتولوژی ژنی به دست آورد. در مقایسه الگوریتم پیشنهادی با دیگر الگوریتم ها، الگوریتمی که تعداد کمپلکس های معنی دار بیشتری را پیش بینی نماید، بهتر است. حد آستانه مقدار میل به همسایگی t در زیر الگوریتم Redundancy – filtering برای کنترل میزان همپوشانی بین هسته های کشف شده استفاده می شود. همان طور که در شکل ۴ نشان داده شد با افزایش مقدار t مقدار F-measure نیز افزایش می یابد. وقتی که t بزرگ تر از ۰/۸ باشد، F-measure در مجموعه داده های DIP ثابت باقی می ماند. در مجموعه داده های MIPS وقتی t بزرگ تر از ۰/۶۵ شود F-measure با مقادیر ثابتی باقی می ماند. در Krogan هم هنگامی که t بزرگ تر 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 مجموعه های داده ای دارای بالاترین مقدار F-measure و ثابت باقی می ماند.

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

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

$$Recall = \frac{|\{R | R \in \mathfrak{R}, \exists K \in \kappa, \alpha(R, K) \geq \theta\}|}{|\mathfrak{R}|} \quad (۱۸)$$

$$F - measure = \frac{2 \times precision \times Recall}{precision + Recall} \quad (۱۹)$$

$$P_value = 1 - \sum_{i=0}^{k-1} \frac{\binom{m}{i} \binom{N-m}{|VP|-i}}{\binom{N}{|VP|}} \quad (۲۰)$$

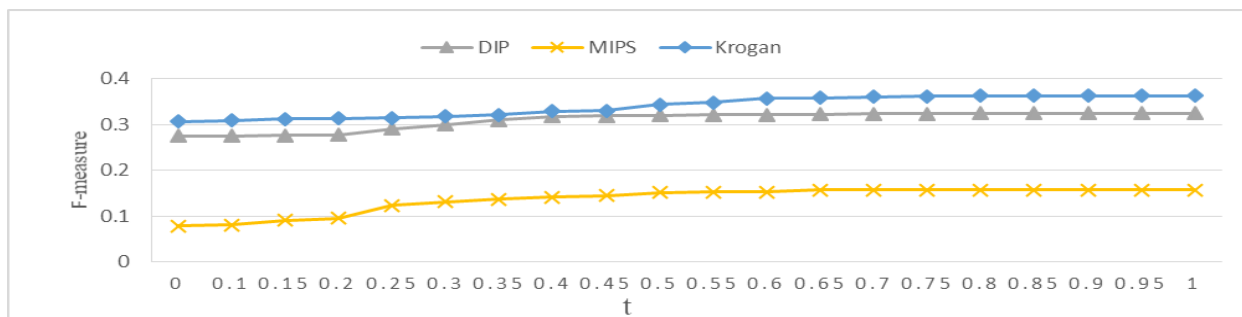
مجموعه $K = \{k_1, k_2, \dots, k_n\}$: مجموعه کمپلکس های کشف شده است.

مجموعه $\mathfrak{R} = \{R_1, R_2, \dots, R_m\}$: مجموعه کمپلکس های واقعی است.

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

θ : حد آستانه همپوشانی بین کمپلکس های واقعی و کمپلکس های کشف شده است. به طور معمول مقدار θ برابر ۰/۵ در نظر گرفته می شود [۳۱].

معیار P_value برای ارزیابی معنی داری بیولوژیکی



شکل ۴: تأثیر مقادیر مختلف حد آستانه میل به همسایگی t و مقادیر F-measure در مجموعه های داده ای

جدول ۱: درصد معنی داری 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 به دست می‌آید.

$$Dave(Gc_v) = \frac{2 \times |E_{Gc_v}| \times W_{ave}}{|V_{Gc_v}|(|V_{Gc_v}| - 1)} \quad (21)$$

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 Precision, Recall, F-measure و P-value عملکرد بالاتری داشته است. به‌طور کلی 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 (در صورتی که برای (Identification of Protein) COACH(Core-Attachment based method) ، [۱۴] ClusteOne(Neighborhood Expansion) و [۱۷] MCODE(Molecular Complex Detection) 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 (در صورتی که مقدار F-measure دیگر روش‌ها [۱۸] CORE ، [۱۳] COACH ، [۱۴] ClusterOne ، [۱۷] MCODE و [۶] به ترتیب برابر ۰/۰۸۷، ۰/۱۴، ۰/۰۹۱ و ۰/۱۰۱ است. برای شبکه Krogan، مقدار Precision و F-measure الگوریتم CEBCOA از

معنی دار هستند و این روش بیشترین نسبت را در میان روش ها به دست آورده است. این نسبت به ترتیب به اندازه ۱۹٪، ۳۰٪، ۸٪، ۱۹٪ و ۱۱٪ بالاتر از روش های IPCA [۱۸]، CORE [۱۳]، COACH [۱۴]، ClusterOne [۱۷] و MCODE [۶] می باشد؛ بنابراین روش پیشنهادی قادر است عملکرد مناسبی در تشخیص کمپلکس های پروتئینی داشته باشد به طوری که معنی بیولوژیکی بهتری نسبت به دیگران حفظ کرده است. در صورتی که روش هایی مانند Core و ClusterOne به علت پیش بینی کمپلکس های کوچک با مقادیر بزرگ P_value، نتایج ضعیفی دارند. کمپلکس های پروتئینی با اندازه بزرگ با احتمال بالایی دارای P_value کوچک تری هستند. الگوریتم های بسیاری برای تشخیص کمپلکس های پروتئینی پیشنهاد شده اند؛ اما هنوز در پیش بینی کمپلکس های پروتئینی از مجموعه داده های مختلف، دقیق و کارآمد نیستند. همچنین موفقیت های هر کدام از روش ها تا حد زیادی بستگی به تکنیک های آزمایشگاهی زیست شناسان دارد که مجموعه داده ای بیولوژیکی قابل اطمینان را برای محاسبات فراهم نمایند؛ بنابراین وقتی دانشمندان کامپیوتر و زیست شناسان با هم کار می کنند برای دانشمندان کامپیوتر با اضافه شدن داده هایی که زیست شناسان فراهم می کنند، راه هایی مطمئن و کافی برای کاویدن دانش جدید از PPI فراهم می شود. در آینده می توان با استفاده از توابع مشابهت همسایگی، برهمکنش های منفی کاذب را پیش بینی کرده تا با حذف آن ها شبکه هایی کامل تر ایجاد کرد. آنگاه با اطلاعات اضافی مانند اطلاعات دامنه ای و داده های بیان ژنی کمپلکس های پروتئینی را با دقت بیشتری پیش بینی کرده و با تحقیق معنی داری بیولوژیکی کمپلکس ها، درک کامل تری از ویژگی های آن ها کشف شود.

با توجه به نقش کمپلکس های پروتئینی در فرایندهای زیستی، شناسایی کمپلکس های پروتئینی یک مسئله مهم در محاسبات زیستی است. در مطالعه حاضر روشی جدید مبتنی بر هسته و پروتئین های الحاقی برای تشخیص کمپلکس های پروتئینی در شبکه های وزن دار ارائه شد. اندازه شباهت معنایی بین جفت پروتئین ها برای تخمین اطمینان برهمکنش های پروتئینی بر اساس آنتولوژی ژنی استفاده شد. با استفاده از اندازه های ترکیبی مرکزیت و ویژگی های زیستی، اهمیت هر کدام از پروتئین ها به دست آمد. آنگاه با مرتب سازی نزولی اندازه ها، پروتئین های مرکزی Seed تشخیص داده شد و سپس با فیلتر پروتئین های نویزی از پروتئین های اساسی، دقت انتخاب پروتئین های مرکزی در مراکز هسته بیشتر شد. در ادامه هسته

سایر الگوریتم ها بیشتر بوده که این مقادیر برای CEBCOA به ترتیب برابر با ۰/۴۲ و ۰/۳۶۳ است، در حالی که برای روش های IPCA [۱۸]، CORE [۱۳]، COACH [۱۴]، ClusterOne [۱۷] و MCODE [۶] مقادیر Precision و F-measure به ترتیب برابر با (۰/۲۹، ۰/۳۰۴)، (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 با (۰/۳۲، ۰/۳۲، ۰/۳۱، ۰/۳۴، ۰/۳۴ و ۰/۱۳) است که حدود ۰/۰۲ کمتر از الگوریتم های IPCA، ClusterOne و COACH و برابر با الگوریتم IPCA می باشد. با توجه به انتخاب دقیق تر مراکز هسته ها به عنوان Seed از طریق پروتئین های اساسی، انتخاب پروتئین هایی با ارتباط عملکردی بهتر در هسته ها، به کارگیری چگالی وزنی قابل قبول در اتصال پروتئین های الحاقی به هسته ها و استفاده از آنتولوژی ژنی، باعث شده که روش پیشنهادی قادر به فیلتر کمپلکس های کم اطمینان و شناسایی کمپلکس های پروتئینی با معنی داری زیستی بالا شود؛ بنابراین هسته هایی که به وسیله CEBCOA شناسایی می شوند می بایستی بعضی عملکردهای مشترک را به اشتراک بگذارند و این مطابقت بیشتری با تعریف هسته کمپلکس ها دارد و همین مسئله باعث پیدا کردن کمپلکس های دقیق تر می شود در صورتی که در روش های دیگر عدم استفاده از مراکز هسته ای دقیق و کشف هسته های اولیه نادقیق، پیش بینی کمپلکس های بی معنی زیستی را بیشتر کرده است و باعث شده که آن ها دارای F-measure پایین تری 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. در صورتی که مقدار F-measure روش های IPCA [۱۸]، CORE [۱۳]، COACH [۱۴]، ClusterOne [۱۷] و MCODE [۶] به ترتیب برابر با (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، ۰/۱۱ کمتر و بهتر از بقیه روش های مطرح شده است. درصد معنی داری کمپلکس های پیش بینی شده در روش های مختلف می تواند به عنوان مقدار کارآمدی آن روش مورد استفاده قرار گیرد. با توجه به نتایج جدول ۱، کمپلکس های پیش بینی شده در روش پیشنهادی ۸۸/۴۸٪

در روش پیشنهادی از درصد معنی‌داری بیولوژیکی بیشتری برخوردار هستند.

تعارض منافع

این مقاله دارای تعارض منافع نمی‌باشد.

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

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* 2014;15:186.
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* 2016;17:371.
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, Berenguères 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 approach for analyzing the functional dissimilarity of gene networks. In 6th 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, Berenguères 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, Waegle 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.

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

Elahi Abdolkarim¹, 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