

عنوان تحقیق:

پیش‌بینی پیوند در شبکه‌ی تعاملات پروتئین-پروتئین
با استفاده از یادگیری ماشین

پژوهشگران:

سحر کیانیان، استادیار دانشگاه شهید رجایی

سعید فرضی، استادیار دانشگاه خواجه نصیر

چکیده:

شبکه‌های پروتئینی اهمیت زیادی در فعالیت‌های بیولوژیکی دارند. تعاملات پروتئین-پروتئین زمانی اتفاق می‌افتد که دو یا چند پروتئین برای انجام برخی فعالیت‌های بیولوژیکی با هم تعامل داشته باشند. به عنوان مثال، سیگنال‌های بیرونی سلول از طریق این فعل و انفعالات به داخل سلول منتقل می‌شوند. شناسایی این تداخلات در شناخت بیماری‌های پیچیده و همچنین طراحی داروها اهمیت زیادی دارد. با در دسترس بودن داده‌های بیولوژیکی عظیم، زیست‌شناسی محاسباتی در موقعیتی قرار دارد که می‌تواند فعل و انفعالات پروتئینی از دست رفته را پیش‌بینی کند. از طرفی، تشخیص زودهنگام و پیگیری درمانی بیماری‌ها به خصوص بیماری‌هایی چون سرطان از اهمیت بالایی برخوردارند تا به این وسیله بتوان مانع آسیب دائمی به سیستم ایمنی و عصبی بدن و یا حتی مرگ این بیماران شد. بنابراین تحقیق جاری در تلاش است تا از طریق شناسایی پیوند جدید بین ژن‌ها در شبکه‌ی تعاملات پروتئین-پروتئین و در نتیجه کمک به تشخیص و درمان به موقع بیماران در جلوگیری از عوارض بیماری‌زا یا حداقل رسانی آن‌ها موثر باشد.

فهرست مطالب

فهرست مطالب	۳
فصل اول: مقدمه	۴
۱-۱ بیان مساله تحقیق	۴
۲-۱ اهداف تحقیق	۵
۳-۱ سوالات یا فرضیه‌های تخصصی	۶
۴-۱ ضرورت و اهمیت تحقیق	۶
۵-۱ کاربرد نتایج تحقیق	۷
فصل ۲- پیشینه تحقیق	۸
۱-۲ مقدمه	۸
۲-۲ جمع بندی	۱۰
منابع و مأخذ:	۱۱

فصل اول: مقدمه

۱-۱ بیان مساله تحقیق

با وجود تحقیقات بسیاری که در زمینه‌ی کشف ژن‌های مرتبط با بیماری‌ها انجام شده است، همچنان شناسایی بخش زیادی از آن‌ها ناشناخته مانده است. اهمیت یافتن ژن‌ها در پیش‌بینی بیماری، بهبود روند درمان و توسعه‌ی دارویی است. یکی از راه‌های موثر در کشف ژن‌های بیماری، تکنیک پیش‌بینی پیوند است که در شبکه‌های پروتئینی انجام می‌شود.

مسئله‌ی پیش‌بینی پیوند گمشده در شبکه‌های پیچیده سال‌هاست که توجه پژوهشگران را به خود جلب کرده است و در کاربردهای فراوانی موثر واقع شده است. برای برخی از شبکه‌ها، به‌ویژه شبکه‌های بیولوژیکی مانند شبکه‌های تعاملات پروتئین-پروتئین، شبکه‌های متابولیکی و شبکه‌های غذایی، کشف پیوندها در آزمایشگاه یا به صورت میدانی پرهزینه است؛ حال می‌توان به جای بررسی کورکورانه همه‌ی پیوندهای ممکن، پیش‌بینی‌هایی مبتنی بر پیوندهای مشاهده‌شده و تمرکز بر پیوندهایی که به احتمال زیاد وجود دارند، را دنبال کرد. اگر پیش‌بینی‌ها به اندازه‌ی کافی دقیق باشند، می‌توانند هزینه‌های آزمایشی را به شدت کاهش دهند.

بطور کلی شبکه‌های تعاملات پروتئین-پروتئین؛ شبکه‌ای از مجموعه‌های پروتئینی هستند که توسط رویدادهای بیوشیمیایی و یا نیروهای الکترواستاتیکی احاطه شده‌اند و عملکرد بیولوژیکی مشخصی را به عنوان یک مجموعه انجام می‌دهند، همچنین این تعاملات پروتئینی، مجموعه کاملی از تعاملات پروتئین-پروتئین یک سیستم بیولوژیکی را توصیف می‌کند.

فرآیند تشخیص پیوند جدید در شبکه‌های پروتئین-پروتئین نیازمند استفاده از مجموعه داده‌های پروتئینی، تعاملات آن‌ها با یکدیگر، ژن‌های از قبل شناسایی شده‌ی مربوط به بیماری و الگوریتمی با پیچیدگی و زمان اجرای کم است تا بتواند پیوندهای جدید را، با درصد احتمال بالاتری در ارتباطات شبکه‌ای بیابد. در این راستا چهار فاکتور اساسی مطرح است که در ادامه معرفی می‌شوند:

- **تاثیر توپولوژیکی شبکه :** فرآیند تشخیص پیوند در شبکه‌های پروتئین-پروتئین نیاز به تخصیص تاثیر توپولوژیکی شبکه‌های مختلف بر یکدیگر، نحوه‌ی تخصیص وزن مناسب به منابع موجود در آزمایش از نظر اهمیت زیست مولکولی آن از دید زیست‌شناسان در الگوریتم و روش به کار گرفته شده است.
- **هزینه‌ی پیش‌بینی پیوند جدید :** پیش‌بینی پیوند جدید بین پروتئین‌ها در شبکه‌ی تعاملات آن‌ها در آزمایشگاه یا به صورت میدانی پرهزینه است، همچون روش پرهزینه‌ی بررسی کورکورانه همه‌ی

پیوندهای ممکن بین پروتئین‌ها در شبکه‌ی آن‌ها، اما اگر پیش‌بینی‌ها به اندازه‌ی کافی دقیق باشند، می‌توانند هزینه‌های آزمایشی را به شدت کاهش دهند.

- **مجموعه داده‌هایی با ابعاد بالا :** شبکه‌های تعاملات پروتئین-پروتئین معمولاً داده‌هایی پراکنده هستند، که علت آن تا حدودی مربوط به نرخ منفی کاذب بالا است، که مانعی می‌شود تا الگوریتم‌ها بتوانند به اطلاعات همسایه‌ها متکی شوند.

- **توزیع درجه‌ای با انحراف بالا :** شبکه‌های تعاملات پروتئین-پروتئین دارای توزیع درجه‌ای منحرف

هستند، به این معنی که آن‌ها بیش از مقدار مورد انتظار ژن هاب دارند. چنین گره‌هایی، هاب، اغلب می‌توانند عملکرد الگوریتم‌های نظری گراف موجود را کاهش دهند.

این فاکتورها اهمیت روش‌های محاسباتی در تشخیص و پیش‌بینی پیوند را آشکار می‌کند که البته از دید پژوهشگران به دور نمانده است [۱]. با وجود پژوهش‌های بسیاری که در این حوزه صورت گرفته است، دقت بالا در پیش‌بینی پیوند با حداقل زمان اجرا، همچنان به عنوان یک چالش مطرح است. نظر به چالش موجود، پژوهش حاضر به دنبال یافتن یک روش موثر پیش‌بینی پیوند در شبکه‌های پروتئینی با کمک یادگیری ماشین است. یادگیری ماشین به عنوان زیرمجموعه‌ای از هوش مصنوعی، یک مدل ریاضی بر اساس داده‌های آموزش ایجاد می‌کند که می‌تواند برای پیش‌بینی یا تصمیم‌گیری مورد استفاده قرار گیرد.

یکی از روش‌های موثر در یادگیری ماشین، مدل یادگیری ترکیبی است که با ترکیب چند دسته‌بندی از یادگیری ماشین برای بهبود عملکرد پیش‌بینی ساخته شده است. در روش سنتی پیش‌بینی، یک دسته‌بندی منفرد، مانند دسته‌بندی ساده بیز، یا درخت‌های تصمیم، برای ساخت مدل پیش‌بینی بر روی یک مجموعه داده استفاده می‌شود. دسته‌بندی‌های منفرد ممکن است برای پیش‌بینی یک نقص خاص تحت یک شرایط خاص، نقاط ضعفی داشته باشند. به نظر می‌رسد، یادگیری ترکیبی بتواند به نقاط قوت دسته‌بندی‌های متعدد برای کشف بهتر در مجموعه داده دست یابد. از این رو، در تحقیق جاری استفاده از روش‌های ترکیبی در الگوریتم‌های یادگیری ماشین مورد توجه خواهد بود.

۱-۲ اهداف تحقیق

هدف اصلی تحقیق جاری بهبود کیفیت تشخیص و پیش‌بینی پیوندها در شبکه‌های پروتئینی است. به این منظور رویکرد یادگیری ماشین پیشنهاد شده است که با توجه به نقش موثر آن در تحقیقات مختلف، انتظار می‌رود بهبود چشمگیری را در این حوزه به دنبال داشته باشد. یادگیری ماشین با مدل سازی رفتار سیستم می‌تواند

در پیش‌بینی پیوندهای گمشده یا پیوندهای آینده در شبکه‌های پروتئینی موثر واقع شود. تحقیق جاری به منظور بهره‌گیری از مزایای روش‌های مختلف یادگیری ماشین، به دنبال یافتن یک الگوریتم ترکیبی کارا برای پیش‌بینی پیوند است.

۱-۳ سوالات یا فرضیه‌های تخصصی

- ۱- چگونه می‌توان یک روش کاربردی و کارا از ترکیب روش‌های مختلف یادگیری ماشین ایجاد کرد؟
- ۲- آیا ترکیب الگوریتم‌های یادگیری ماشین نسبت به روش‌های پایه برای پیش‌بینی پیوند در شبکه‌های پروتئینی برتری دارد؟
- ۳- چه ویژگی‌هایی در شبکه‌های پروتئینی وجود دارد که آن را در به کارگیری الگوریتم‌های پیش‌بینی پیوند از سایر شبکه‌های پیچیده متمایز می‌کند؟

۱-۴ ضرورت و اهمیت تحقیق

شبکه‌های پروتئینی اهمیت زیادی در فعالیت‌های بیولوژیکی دارند. تعاملات پروتئین-پروتئین زمانی اتفاق می‌افتد که دو یا چند پروتئین برای انجام برخی فعالیت‌های بیولوژیکی با هم تعامل داشته باشند. به عنوان مثال، سیگنال‌های بیرونی سلول از طریق این فعل و انفعالات به داخل سلول منتقل می‌شوند. شناسایی این تداخلات در شناخت بیماری‌های پیچیده و همچنین طراحی داروها اهمیت زیادی دارد. با در دسترس بودن داده‌های بیولوژیکی عظیم، زیست‌شناسی محاسباتی در موقعیتی قرار دارد که می‌تواند فعل و انفعالات پروتئینی از دست رفته را پیش‌بینی کند. از طرفی، تشخیص زودهنگام و پیگیری درمانی بیماری‌ها به خصوص بیماری‌هایی چون سرطان از اهمیت بالایی برخوردارند تا به این وسیله بتوان مانع آسیب دائمی به سیستم ایمنی و عصبی بدن و یا حتی مرگ این بیماران شد. بنابراین تحقیق جاری در تلاش است تا از طریق شناسایی پیوند جدید بین ژن‌ها در شبکه‌ی تعاملات پروتئین-پروتئین و در نتیجه کمک به تشخیص و درمان به موقع بیماران در جلوگیری از عوارض بیماری‌زا یا حداقل رسانی آن‌ها موثر باشد.

۵-۱ کاربرد نتایج تحقیق

- تشخیص زودهنگام بیماری: شناسایی ارتباط miRNA-بیماری می تواند ما را قادر به درک بیماری ها در سطح miRNA کند. بنابراین، آشکار کردن بیشتری از ارتباطات بالقوه miRNA-بیماری یک موضوع حیاتی در حوزه ی زیست پزشکی است.
- توسعه ی دارو: پیش بینی پاسخ دارویی مسئله ای مهم است که در آن از مشخصات مولکولی یک نمونه معین برای پیش بینی اثر یک داروی معین بر آن نمونه استفاده می شود. شبکه های مولکولی زمینه ای کاربردی برای ادغام ویژگی های ژنومی فراهم می کنند و در نتیجه منجر به مدل های پیش بینی قوی و قابل تکرار می شوند. برای غلبه بر این چالش ها، می توانیم پیش بینی پاسخ دارویی را به عنوان یک مشکل پیش بینی پیوند فرموله کنیم.
- پیش بینی بیماری های ژنتیکی: همه ی بیماری ها یک جزء ژنتیکی دارند؛ با این حال، میزان مشارکت ژن ها در ایجاد بیماری متفاوت است و هنوز چیزهای زیادی برای یادگیری باقی مانده است. پیشرفت ها در درک مکانیسم های ژنتیکی پشت این بیماری، توسعه آزمایش های تشخیصی اولیه، درمان های جدید یا مداخلات را برای پیشگیری از شروع بیماری یا به حداقل رساندن شدت بیماری امکان پذیر می سازد. این بخش اطلاعاتی در مورد اهمیت علائم بالینی مانند سابقه خانوادگی که ممکن است حاکی از یک بیماری ژنتیکی باشد، کاربردهای مختلف آزمایش ژنتیکی و انواع مختلف بیماری های ژنتیکی ارائه می دهد.
- پروژه ی ژنوم انسان، نقشه برداری ژنتیکی و شبیه سازی ژن های جهش یافته را که مستعد ابتلا به بسیاری از بیماری های انسانی هستند، آسان تر کرده است. علاوه بر بیماری هایی که به روش ساده ی ذهنی به ارث می رسند، بسیاری از بیماری های دژنراتیو و عفونی و حساسیت ها به توهین های محیطی وجود دارد که در آنها ساختار ژنتیکی یک فرد در روند بیماری نقش دارد. هنگامی که یک ژن یا ژن های مرتبط با بیماری شبیه سازی شد، می توان تشخیص های مبتنی بر DNA را برای تشخیص اشکال تغییر یافته ی ژن که مستعد ابتلا به بیماری اند را طراحی کرد. توانایی پیش بینی توسعه ی بیماری، مداخله ی زودهنگام را برای محدود کردن شدت بیماری یا استفاده از ژن درمانی برای درمان اختلالات ارثی ممکن می سازد.

فصل ۲- پیشینه تحقیق

۲-۱ مقدمه

فرآیندهای بیولوژیکی سلول توسط مولکول‌های زیستی مختلف، به‌ویژه پروتئین‌ها و تعاملات آن‌ها از طریق تعاملات فیزیکی موقت یا طولانی‌مدت انجام می‌شود. بنابراین، بررسی ارتباطات پروتئین-پروتئین ممکن است به درک بهتری در عملکردهای سلولی و مکانیسم‌های بیماری، و همچنین بهبود در طراحی دارو منجر شود [۲].

علیرغم تلاش‌های تجربی استثنایی برای ترسیم تعامل پروتئین‌های انسانی، به دلیل ناقص بودن اطلاعات و داده‌ها، توانایی ما را برای درک ریشه‌های مولکولی بیماری‌های انسانی محدود می‌کند. ابزارهای محاسباتی یک جایگزین برای شناسایی تعاملات پروتئین-پروتئین است که از نظر بیولوژیکی در مسیر شناسایی ژن‌های بیمار کمک می‌کنند. در حالی که روش‌های پیش‌بینی پیوند، پروتئین‌ها را بر اساس شباهت بیولوژیکی مبتنی بر شبکه به یکدیگر متصل می‌کنند، اما این نکته نیز حائز اهمیت است که لزوماً پروتئین‌های در ارتباط با یکدیگر مشابه نیستند و پروتئین‌های مشابه نیز لزوماً برهم‌کنش ندارند [۳].

شباهت گره‌ها را می‌توان با ویژگی‌های اساسی گره‌ها تعریف کرد؛ به عنوان مثال، اگر دو نفر دارای سن، جنس و شغل یکسان باشند، می‌توان گفت که آن‌ها شبیه هم هستند. و برای گروه دیگری شباهت‌ها فقط بر اساس ساختار شبکه است. برخی از شاخص‌های شباهت مانند؛ همسایگان مشترک، ضریب جاکارد، شاخص آدامیک-آدار و پیوست ترجیحی، شاخص‌های وابسته به گره هستند که فقط به اطلاعات مربوط به درجه‌ی گره و نزدیکترین همسایه نیاز دارند. در حالی که شاخص کاتز، زمان ضربه، زمان رفت و آمد، رتبه صفحه ریشه دار و شاخص بلوندل به شاخص‌های وابسته به مسیر تعلق دارند.

در پژوهش استنفیلد و همکاران نشان داده شد که miRNA یکی از مهم‌ترین اجزای سلولی است که در فرآیندهای بیولوژیکی متعددی از جمله رشد سلولی، تکثیر، انتقال سیگنال، تمایز، آپوپتوز، عفونت ویروسی، متابولیسم، نقش حیاتی دارد. شواهد انباشته نشان داده است که جهش‌های miRNA با سرطان‌های مختلف انسانی مرتبط است که نشان می‌دهد miRNA ها می‌توانند به عنوان سرکوب‌کننده‌های تومور و انکوژن عمل کنند. با توجه به رابطه‌ی نزدیک بین miRNA ها و بیماری‌های انسانی، شناسایی ارتباط بالقوه بیشتر بین miRNA ها و بیماری‌ها در حال حاضر یک هدف مهم در حوزه ی زیست پزشکی است زیرا ممکن است کمک زیادی به تحقیقات در مورد بیماری کند. با این حال، اگر همه جفت‌های miRNA-بیماری را با آزمایش

بیولوژیکی آزمایش کنیم، بسیار گران و زمان‌بر خواهد بود. بنابراین، برای جلوگیری از اتلاف پول و زمان، لازم است پتانسیل‌ترین جفت‌های miRNA-بیماری را پیش‌بینی کرد. از آنجایی که پایگاه‌های اطلاعاتی بیشتری در دسترس هستند، توسعه‌ی مدل محاسباتی راهی موثر برای آشکار کردن ارتباط بالقوه بین miRNA ها و بیماری‌ها خواهد بود.

در مقاله‌ی [۴]، با ترکیب مزایای هر دو روش، یک مدل محاسباتی جدید از یادگیری گروهی و پیش‌بینی پیوند برای پیش‌بینی انجمن بیماری (miRNA (ELLPMDA و ارتباط‌های بالقوه miRNA-بیماری ارائه شده است. به جای ترکیب نتایج چند دسته‌بندی، ترکیب وزنی رتبه‌های ارائه‌شده توسط سه الگوریتم کلاسیک مبتنی بر شباهت، همسایگان مشترک، شاخص جاکارد و شاخص کاتز را خروجی می‌دهد.

داده‌های شبکه‌ی تعاملات پروتئین-پروتئین را می‌توان توسط شبکه‌هایی مدل‌سازی کرد، جایی که گره‌ها در شبکه‌ها پروتئین‌ها را نشان می‌دهند و یال‌های بین گره‌ها نشان‌دهنده تعامل فیزیکی بین پروتئین‌ها هستند. این شبکه‌ها همراه با سایر پروتئین‌های داده‌های ژنومیک عملکردی، فرصت‌های بی‌سابقه‌ای را برای دانشمندان بیولوژیکی و محاسباتی فراهم می‌کند تا سلول را در سطح سیستمی درک کنند؛ به عنوان مثال، تجزیه و تحلیل جهانی شبکه‌های تعاملات پروتئین-پروتئین، ارتباطات مهمی را بین توپولوژی و عملکرد نشان داده است. شبکه‌های تعاملات پروتئین-پروتئین همچنین برای پیش‌بینی عملکرد ژن، مسیرهای عملکردی یا کمپلکس‌های پروتئینی، با روش‌های نظارت شده و بدون نظارت، استفاده شده‌اند [۱].

علاوه بر این، اخیراً تلاش زیادی برای ترکیب شبکه‌های تعاملات پروتئین-پروتئین برای به دست آوردن درک مکانیکی بهتر از بیماری‌های پیچیده و بهبود، تشخیص و درمان بیماری‌ها انجام شده است. با این حال، اندازه و پیچیدگی روزافزون شبکه‌های تعاملات پروتئین-پروتئین چالش‌های متعددی را برای زیست‌شناسان ایجاد می‌کند. اول، شبکه‌های تعاملات پروتئین-پروتئین اغلب دارای نرخ مثبت کاذب بالا و نرخ منفی کاذب حتی بالاتری هستند.

دوماً، شبکه‌های تعاملات پروتئین-پروتئین معمولاً پراکنده هستند، تا حدی که دلیل نرخ بالای منفی کاذب، که مانعی برای الگوریتم‌هایی ایجاد می‌کند که به اطلاعات همسایه‌ها متکی هستند، به عنوان مثال، در پیش‌بینی عملکرد ژن.

ثالثاً، شبکه‌های تعاملات پروتئین-پروتئین دارای توزیع درجه‌ای منحرف هستند، به این معنی که آنها بیش از مقدار مورد انتظار ژن‌ها دارند. چنین گره‌های هاب اغلب می‌توانند عملکرد الگوریتم‌های نظری گراف

موجود را کاهش دهند (مثلاً برای پیش‌بینی کمپلکس‌های پروتئینی) که اغلب برای شبکه‌هایی با توزیع درجه نسبتاً یکنواخت طراحی شده‌اند [۱].

در سال‌های اخیر، اختلال در تنظیم miRNA ها به عنوان دلیل اصلی رفتار ناهنجار سلولی توسط بسیاری از مطالعات تایید شده است. علاوه بر این، پلی‌مورفیسم‌های تک نوکلئوتیدی واقع در مکان‌های اتصال به miRNA احتمالاً بر بیان هدف miRNA تأثیر می‌گذارند و ممکن است در حساسیت انسان به بیماری‌های رایج نقش داشته باشند.

علاوه بر این، کاهش miR-101 در بیان بیش از حد سیکلواکسیژناز-۲ در سلول‌های سرطانی روده بزرگ انسان نقش دارد. همچنین ارتباط miR-214 با سرطان تخمدان انسان از طریق القای بقای سلولی و مقاومت به سیس پلاتین تایید شده است. مثال دیگر این است که مطالعات بیولوژیکی نشان می‌دهد miR-34a با سرکوب مستقیم CD44 سلول‌های بنیادی سرطان پروستات و متاستاز را مهار می‌کند. علاوه بر این، تحقیقات اشاره کرده بودند که خوشه 1-miR-16-miR-15a با هدف قرار دادن چندین فعالیت انکوژنیک، سرطان پروستات را کنترل می‌کند.

۲-۲ جمع بندی

از بررسی کارهای موجود می‌توان دریافت که بیشترین مقالات از مجموعه داده‌ی STRING استفاده کرده بودند و تعدادی از آن‌ها با کمک اعضای تیم به گردآوری اطلاعات مورد نیاز در این زمینه پرداخته بودند. همچنین الگوریتم‌های به کار رفته در این مقالات بیشتر به صورت پیاده‌روی تصادفی یا زیرمجموعه‌ای از آن پیاده‌سازی شده است. نکته‌ی حائز اهمیت این است که مقالات سعی داشته‌اند تا نتایج تحقیق خود را با بخشی از سایر روش‌هایی که قبل از آن تاریخ منتشر شده بود، مقایسه کنند و نتایج معتبری را عرصه نمایند که علاوه بر معیارهایی که برای سنجش و ارزیابی دقت و صحت الگوریتم به کار گرفته می‌شوند، مقایسه‌ای هم با دیگر الگوریتم‌ها و نتایج علمی ثابت شده، داشته باشند.

- [1] Chen, S.J., Liao, D.L., Chen, C.H., Wang, T.Y. and Chen, K.C., 2019. Construction and analysis of protein-protein interaction network of heroin use disorder. *Scientific reports*, 9(1), pp.1-9.
- [2] Balogh, O.M., Benczik, B., Horváth, A., Pétervári, M., Csermely, P., Ferdinandy, P. and Ágg, B., 2022. Efficient link prediction in the protein–protein interaction network using topological information in a generative adversarial network machine learning model. *BMC bioinformatics*, 23(1), pp.1-19.
- [3] Berahmand, K., Nasiri, E., Forouzandeh, S. and Li, Y., 2022. A preference random walk algorithm for link prediction through mutual influence nodes in complex networks. *Journal of king saud university-computer and information sciences*, 34(8), pp.5375-5387..
- [4] Chen, X., Zhou, Z. and Zhao, Y., 2018. ELLPMDA: ensemble learning and link prediction for miRNA-disease association prediction. *RNA biology*, 15(6), pp.807-818.
- [5] Stanfield, Z., Coşkun, M. and Koyutürk, M., 2017. Drug response prediction as a link prediction problem. *Scientific reports*, 7(1), pp.1-13.
- [6] Yu, L., Zhou, D., Gao, L. and Zha, Y., 2021. Prediction of drug response in multilayer networks based on fusion of multiomics data. *Methods*, 192, pp.85-92.
- [7] Kurakar, M. and Izudheen, S., 2014. Link prediction in protein-protein networks: survey. *IJCTT*, pp.164-168.
- [8] Lei, C. and Ruan, J., 2013. A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity. *Bioinformatics*, 29(3), pp.355-364.
- [9] Kovács, I.A., Luck, K., Spirohn, K., Wang, Y., Pollis, C., Schlabach, S., Bian, W., Kim, D.K., Kishore, N., Hao, T. and Calderwood, M.A., 2019. Network-based prediction of protein interactions. *Nature communications*, 10(1), pp.1-8.
- [10] Iakovidou, N., Symeonidis, P. and Manolopoulos, Y., 2010, November. Multiway spectral clustering link prediction in protein-protein interaction networks. In *Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine* (pp. 1-4). IEEE.
- [11] Nasiri, E., Berahmand, K., Rostami, M. and Dabiri, M., 2021. A novel link prediction algorithm for protein-protein interaction networks by attributed graph embedding. *Computers in Biology and Medicine*, 137, p.104772.

- [12] Wang, X., Gulbahce, N. and Yu, H., 2011. Network-based methods for human disease gene prediction. *Briefings in functional genomics*, 10(5), pp.280-293.
- [13] Lee, I. and Nam, H., 2018. Identification of drug-target interaction by a random walk with restart method on an interactome network. *BMC bioinformatics*, 19(8), pp.9-18.
- [14] Zhang, Y., Dai, L., Liu, Y., Zhang, Y. and Wang, S., 2017. Identifying novel fruit-related genes in *Arabidopsis thaliana* based on the random walk with restart algorithm. *PLoS One*, 12(5), p.e0177017.
- [15] Li, J., Wang, L., Guo, M., Zhang, R., Dai, Q., Liu, X., Wang, C., Teng, Z., Xuan, P. and Zhang, M., 2015. Mining disease genes using integrated protein–protein interaction and gene–gene co-regulation information. *FEBS open bio*, 5, pp.251-256.
- [16] Weiping Liu & Linyuan Lu. (2010). Link prediction based on local random walk. *EPL Journal*, 58007-p1- 58007-p6.
- [17] Luo P, Tian LP, Ruan J, Wu FX. Disease Gene Prediction by Integrating PPI Networks, Clinical RNA-Seq Data and OMIM Data. *IEEE/ACM Trans Comput Biol Bioinform*. 2019 Jan-Feb;16(1):222-232. Epub 2017 Nov 7. PMID: 29990218.
- [18] Chen, S.J., Liao, D.L., Chen, C.H., Wang, T.Y. and Chen, K.C., 2019. Construction and analysis of protein-protein interaction network of heroin use disorder. *Scientific reports*, 9(1), pp.1-9.
- [19] A. Papadimitriou, P. Symeonidis and Y. Manolopoulos, "Scalable Link Prediction in Social Networks Based on Local Graph Characteristics," 2012 Ninth International Conference on Information Technology - New Generations, 2012, pp. 738-743.
- [20] Shabaz, M. and Garg, U., 2021. Predicting future diseases based on existing health status using link prediction. *World Journal of Engineering*.
- [21] Cui, X., Shen, K., Xie, Z., Liu, T. and Zhang, H., 2017. Identification of key genes in colorectal cancer using random walk with restart. *Molecular medicine reports*, 15(2), pp.867-872.
- [22] Iakovidou, N., Symeonidis, P. and Manolopoulos, Y., 2010, November. Multiway spectral clustering link prediction in protein-protein interaction networks. In *Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine* (pp. 1-4). IEEE.

- [23] Xia, F., Liu, J., Nie, H., Fu, Y., Wan, L. and Kong, X., 2019. Random walks: A review of algorithms and applications. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 4(2), pp.95-107.
- [24] Sen, S.C., Kasim, S., Fudzee, M.F.M., Abdullah, R. and Atan, R., 2017. Random walk from different perspective. *Acta Electronica Malaysia*, 1(2), pp.26-27.
- [25] Martínez, V., Berzal, F. and Cubero, J.C., 2016. A survey of link prediction in complex networks. *ACM computing surveys (CSUR)*, 49(4), pp.1-33.
- [26] Mirzaei, A., Pourahmadi, V., Soltani, M. and Sheikhzadeh, H., 2020. Deep feature selection using a teacher-student network. *Neurocomputing*, 383, pp.396-408
- [27] Chen, Y., Wang, W., Liu, J., Feng, J. and Gong, X., 2020. Protein interface complementarity and gene duplication improve link prediction of protein-protein interaction network. *Frontiers in Genetics*, 11, p.291.

Abstract :

Protein networks are very important in biological activities. Protein-protein interactions occur when two or more proteins interact to perform some biological activity. For example, signals from outside the cell are transmitted into the cell through these interactions. Identifying these interactions is very important in understanding complex diseases and designing drugs. With the availability of massive biological data, computational biology is in a position to predict missing protein interactions. On the other hand, early diagnosis and treatment follow-up of diseases, especially diseases such as cancer, are of great importance in order to prevent permanent damage to the immune and nervous system of the body or even the death of these patients. Therefore, current research is trying to be effective in preventing or minimizing pathogenic complications through the identification of new links between genes in the network of protein-protein interactions and thus helping to diagnose and treat patients in time.

Title:

**Link prediction in protein-protein interaction network using
machine learning**

Researchers:

**Sahar kianian, Shahid Rajaee University,
sahar.kianian@sru.ac.ir**

**Saeed Farzi, K. N. Toosi University of technology,
saeedfarzi@kntu.ac.ir**