



دانشکده علوم رایانه و فناوری اطلاعات

مهندسی کامپیوتر - هوش مصنوعی

کشف ماژول‌ها در شبکه‌های پروتئین - پروتئین با رویکردهای شبکه‌های عصبی گرافی

پایان‌نامه‌ی کارشناسی ارشد

سمانه طجرلو

استاد راهنما: زهرا نریمانی

۱۷ دی ۱۴۰۴

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

تقدیم به آنهایی که می‌خوانند بیشتر بدانند.

شکر و قدردانی

در اینجا از همه دوستانم که در این سال ها به من کمک کرده اند تشکر می کنم.

چکیده

بیوانفورماتیک یک حوزه میان‌رشته‌ای است که با استفاده از علوم زیست‌شناسی، کامپیوتر، ریاضیات و آمار به ذخیره‌سازی و تحلیل داده‌های زیستی می‌پردازد. با پایان‌یافتن پروژه توالی‌یابی ژنوم انسان و ورود به دوره‌ی پساژنی، تحقیقات پروتئومیک به یکی از مهم‌ترین حوزه‌های علوم زیستی تبدیل شده است. پروتئومیک به مطالعه ویژگی‌های پروتئین‌ها برای توصیف ساختار، عملکرد و کنترل سیستم‌های زیستی می‌پردازد. پروتئین‌ها اغلب به تنهایی عمل نمی‌کنند، بلکه با هم تعامل دارند و برای انجام وظایف زیستی، به مولکول‌های بزرگ‌تری تبدیل می‌شوند. تعاملات بین پروتئین‌ها را به کمک ساختار شبکه‌ای به نام شبکه تعامل پروتئین-پروتئین نمایش می‌دهند. یک ترکیب پروتئینی در شبکه‌های PPI یک ساختار مولکولی است که هم از نظر ویژگی و هم از نظر ساختاری از پروتئین‌های سازگار با هم تشکیل شده است. با تحلیل شبکه‌های PPI می‌توانیم این مجموعه از پروتئین‌ها را شناسایی کنیم. یکی از مسائل مهم در بیوانفورماتیک کشف ماژول‌های پروتئین در شبکه‌های تعامل پروتئین-پروتئین است. کشف این ماژول‌ها معادل مسئله‌ی کشف انجمن در گراف است. در بسیاری از کاربردهای بیوانفورماتیکی کشف ماژول‌های پروتئینی با استفاده از الگوریتم‌های کشف انجمن در گراف انجام می‌شود. در این پژوهش ما قصد داریم روشی ویژه برای کشف انجمن در شبکه‌های تعامل پروتئینی طراحی کنیم که علاوه بر در نظر گرفتن ساختار گرافی برای شناسایی ماژول‌ها به ویژگی‌های زیستی پروتئین‌ها نیز توجه دارد. برای مثال، استفاده از اطلاعات زیستی پروتئین‌ها که در پایگاه‌های داده‌ای مانند GO و KEGG ذخیره شده‌اند، همراه با داده‌های بیان ژنی و ترکیب این اطلاعات با شبکه PPI می‌تواند به شناسایی دقیق‌تر و کارآمدتر ماژول‌های پروتئینی کمک کند. از این روی، در این پژوهش ما قصد معرفی یک الگوریتم خوشه‌بندی برای شبکه‌های PPI بر پایه شبکه‌های عصبی گرافی و با در نظر گرفتن ویژگی‌های گره‌ها داریم.

واژه‌های کلیدی: شبکه‌های عصبی گراف‌ی، تعامل پروتئین-پروتئین، شناسایی مازول‌های عملکردی،
خوشه‌بندی گراف‌های دارای ویژگی

فهرست مطالب

چکیده	پنج
پیش‌گفتار	۱
۱ معرفی پژوهش	۲
۱.۱ مقدمه	۲
۲.۱ بیان مسئله	۳
۳.۱ اهمیت و ضرورت انجام پژوهش	۶
۴.۱ پرسش‌های پژوهش	۷
۵.۱ روش پژوهش	۷
۶.۱ جمع‌بندی	۷
۲ مفاهیم بنیادی	۸
۱.۲ مقدمه	۸
۲.۲ مفاهیم زیستی	۹
۱.۲.۲ پروتئین	۹
۲.۲.۲ ماژول‌های عملکردی	۱۰
۳.۲.۲ بیان ژن	۱۱

۴.۲.۲	پایگاه داده هستی‌شناسی ژن	۱۲
۵.۲.۲	شبکه‌های PPI و ویژگی‌های آنها	۱۴
۳.۲	مفاهیم محاسباتی	۱۵
۱.۳.۲	یادگیری ماشین	۱۵
۲.۳.۲	یادگیری عمیق	۱۶
۳.۳.۲	یادگیری نظارت شده	۱۷
۴.۳.۲	یادگیری بدون نظارت	۱۷
۵.۳.۲	گراف	۱۷
۶.۳.۲	شبکه‌های عصبی گرافی	۱۸
۷.۳.۲	شبکه‌های عصبی گرافی پیچشی	۲۰
۸.۳.۲	شبکه‌های عصبی گرافی توجه محور	۲۱
۹.۳.۲	شبکه‌های عصبی گرافی با انتقال دانش بین لایه‌ها	۲۳
۱۰.۳.۲	تعبیه گره‌ها به روش Node2Vec	۲۵
۱۱.۳.۲	خوشه‌بندی گراف با گره‌های ویژگی‌دار	۲۶
۱۲.۳.۲	دسته‌بندی و روش‌های کلی خوشه‌بندی گراف	۲۷
۴.۲	معیارهای ارزیابی	۲۸
۱.۴.۲	شباهت همسایگی	۲۸
۲.۴.۲	دقت	۳۰
۳.۴.۲	بازیابی	۳۰
۴.۴.۲	امتیاز F	۳۰
۵.۴.۲	صحت	۳۱
۳۲	بررسی منابع	۳۲

۳۲	خوشه‌بندی گراف با گره‌های ویژگی‌دار	۱.۳
۳۹	پیش‌بینی مجموعه‌های پروتئینی	۲.۳
۳۹	روش‌های بر پایه شبکه	۱.۲.۳
۴۰	روش‌های مبتنی بر آگاهی از زمینه‌های زیستی	۲.۲.۳
۴۲	روش‌های مبتنی بر شبکه‌های عصبی گرافی	۳.۲.۳
۴۵	روش شناسی پژوهش	۴
۴۵	مقدمه	۱.۴
۴۶	مجموعه داده	۲.۴
۴۷	مجموعه داده شبکه‌های PPI	۱.۲.۴
۴۹	استخراج ویژگی‌ها از پایگاه هستی‌شناسی ژن	۲.۲.۴
۵۱	مجموعه‌های پروتئینی مرجع و پروتکل ارزیابی	۳.۲.۴
۵۲	روش پیشنهادی	۳.۴
۶۶	واژه‌نامه انگلیسی به فارسی	
۶۷	واژه‌نامه فارسی به انگلیسی	

پیش‌گفتار

در پژوهش حاضر اقدام به معرفی یک روش به منظور شناسایی مجموعه‌های عملکردی در شبکه‌های تعامل پروتئین-پروتئین به کمک شبکه‌های عصبی گرافی کرده‌ایم. در طول این پژوهش مطالعه و فهم مفاهیم زیستی یکی از چالش‌های اصلی این تحقیق بوده است. همچنین توسعه یک روش جدید برپایه شبکه‌های عصبی گرافی نیازمند فهم عمیق از نحوه عملکرد این شبکه‌ها است. مدل پیشنهادی عملکرد بهتری نسبت به روش‌های موجود نشان داده و امکان پژوهش و بررسی این روش بر روی سایر مجموعه داده‌ها (به جز شبکه‌های پروتئین-پروتئین) می‌تواند مورد مطالعه قرار بگیرد.

فصل ۱

معرفی پژوهش

۱.۱ مقدمه

پروتئین‌ها مولکول‌های بزرگ و پیچیده‌ای هستند که وظایف زیستی حیاتی‌ای مانند نقش ساختاری و حمایتی از سلول‌ها، عملکرد پادتنی، نقش‌های پیام‌رسانی و یا نقش آنزیمی را عهده دار هستند. بسیاری از فرآیندهای زیستی به وسیله مجموعه‌ای از پروتئین‌ها انجام می‌گیرد که ماژول‌های عملکردی نامیده می‌شوند. شناخت هر چه بهتر این مجموعه‌های پروتئینی به درک بهتر ما از فرآیندهای زیستی و همچنین درک نقش پروتئین‌های کمتر شناخته شده کمک شایانی می‌کند. تشخیص این ماژول‌ها به کمک روش‌های آزمایشگاهی سخت و هزینه‌بر است از این روی تلاش‌های بسیاری (مقاله سایت بزنیم) در جهت ارائه روش‌های محاسباتی کارا به منظور شناسایی این مجموعه‌های پروتئینی انجام شده است. در ادامه این بخش به بیان بهتر مسئله پیش‌رو و همچنین مفاهیم بنیادی مورد نیاز می‌پردازیم.

زیست‌انفورماتیک سعی بر پاسخ به مسائل و پرسش‌های زیست‌شناسی به کمک ابزارهای محاسباتی و مدل‌سازی‌های آماری دارد. یافتن پاسخ مناسب برای هر یک از این مسائل می‌تواند تأثیر به‌سزایی در

فهم بیشتر ما از عملکردهای زیستی داشته باشد. در این پژوهش، ما بر روی پروتئین‌ها تمرکز کرده‌ایم و هدف شناسایی مجموعه‌های پروتئینی، به کمک شبکه برهم‌کنش پروتئین-پروتئین می‌باشد.

۲.۱ بیان مسئله

با پایان یافتن پروژه توالی‌یابی ژنوم انسان و ورود به دوره‌ی پساژنی^۱، پژوهش‌های پروتئومیک^۲ به یکی از مهم‌ترین و فعال‌ترین حوزه‌های علوم زیستی تبدیل شده‌اند. پروتئومیک به مطالعه جامع ویژگی‌های پروتئین‌ها با هدف توصیف ساختار، عملکرد و سازوکارهای کنترلی سیستم‌های زیستی می‌پردازد. پروتئین‌ها عموماً به صورت منفرد عمل نمی‌کنند، بلکه از طریق تعامل با یکدیگر مجموعه‌هایی را تشکیل می‌دهند که انجام بسیاری از وظایف حیاتی سلولی را امکان‌پذیر می‌سازند. تعاملات پروتئین-پروتئین^۳ نقشی اساسی در فرآیندهایی نظیر تکثیر ماده ژنتیکی^۴، کنترل بیان ژن^۵، انتقال سیگنال‌های سلولی^۶ و مرگ برنامه‌ریزی‌شده سلولی^۷ ایفا می‌کنند. از این‌رو، تحلیل شبکه‌های PPI برای درک عمیق‌تر سازماندهی و عملکرد سلولی امری ضروری محسوب می‌شود [۱].

زیست‌انفورماتیک^۸ به عنوان یک حوزه میان‌رشته‌ای، با تلفیق دانش زیست‌شناسی، علوم کامپیوتر، ریاضیات و آمار، به ذخیره‌سازی، مدیریت و تحلیل داده‌های زیستی می‌پردازد. این علم با بهره‌گیری از ابزارها و فناوری‌های محاسباتی، داده‌های مرتبط با توالی‌های دی‌ان‌ای^۹، آران‌ای^{۱۰} و پروتئین‌ها را پردازش و تفسیر می‌کند و با توجه به حجم عظیم داده‌ها و اهمیت استخراج دانش کاربردی از آن‌ها،

¹ Postgenomic era

² Proteomics

³ Protein-protein interactions (PPI)

⁴ Gene substance copy

⁵ Gene expression control

⁶ Cellular signal transduction

⁷ Cell apoptosis

⁸ Bioinformatic

⁹ DNA

¹⁰ RNA

جایگاهی کلیدی در پژوهش‌های نوین زیستی یافته است [۲].

مطالعات زیستی نشان می‌دهد که یک مجموعه پروتئینی^۱ در شبکه‌های PPI به صورت یک ساختار مولکولی منسجم تعریف می‌شود که از پروتئین‌هایی با سازگاری عملکردی و ساختاری تشکیل شده است [۱]. به بیان دیگر، پروتئین‌هایی که در شبکه PPI با یکدیگر تعامل دارند، غالباً از منظر کارکردهای زیستی نیز شباهت‌های معناداری از خود نشان می‌دهند. بر این اساس، زیرشبکه‌های به هم پیوسته و با تراکم بالای پروتئین‌ها می‌توانند به عنوان ماژول‌های عملکردی^۲ یا مجموعه‌های پروتئینی در نظر گرفته شوند که در انجام فرآیندهای زیستی خاص نقش دارند [۳]. شناسایی این ساختارها علاوه بر امکان بررسی تعامل میان مجموعه‌های پروتئینی مختلف، می‌تواند به کشف مجموعه‌های پروتئینی ناشناخته نیز منجر شود [۱].

یکی از رویکردهای کارآمد برای مطالعه شبکه‌های PPI، مدل‌سازی و تحلیل آن‌ها از منظر نظریه گراف و شبکه‌های پیچیده است. با اتخاذ این دیدگاه و با توجه به مقیاس بزرگ داده‌های زیستی، یکی از چالش‌های اساسی در دوره پساژنی، طراحی الگوریتم‌های بهینه و مقیاس‌پذیر به منظور شناسایی مؤثر ماژول‌های عملکردی و مجموعه‌های پروتئینی زیستی است [۴]. از آنجایی که پروتئین‌های یک مجموعه پروتئینی در شبکه PPI دارای تعاملات متراکم و فراوانی با یکدیگر هستند، این نواحی متراکم در شبکه را می‌توان به عنوان مجموعه‌های پروتئینی احتمالی در نظر گرفت. در نتیجه، مسئله شناسایی مجموعه‌های پروتئینی شباهت زیادی به مسئله خوشه‌بندی در شبکه‌های پیچیده دارد [۵] و می‌توان آن را به صورت یک مسئله خوشه‌بندی گرافی مدل‌سازی کرد.

بخش عمده‌ای از پژوهش‌های پیشین در این حوزه صرفاً بر پایه اطلاعات ساختاری شبکه‌های PPI ارائه شده‌اند [۶]، [۷]، در حالی که امروزه داده‌های غنی و متنوعی برای توصیف ویژگی‌های پروتئین‌ها در دسترس است. به عنوان نمونه، پایگاه داده هستی‌شناسی ژن^۳ اطلاعات جامعی درباره ژن‌ها و پروتئین‌ها

¹ Protein complex

² Functional module

³ Gene Ontology

از منظر فرآیندهای زیستی، عملکردهای مولکولی و مؤلفه‌های سلولی فراهم می‌کند و به طور گسترده مورد استفاده پژوهشگران قرار گرفته است [۸]. افزون بر این، از آنجایی که پروتئین‌ها محصولات بیان ژن هستند، برخی مطالعات از داده‌های بیان ژنی مرتبط با هر پروتئین نیز به منظور توصیف دقیق‌تر آن‌ها در شبکه‌های PPI بهره برده‌اند [۹]. ترکیب این اطلاعات تکمیلی با ساختار شبکه PPI می‌تواند به شناسایی دقیق‌تر و مؤثرتر مجموعه‌های پروتئینی منجر شود.

برای بهره‌برداری مناسب از این اطلاعات، نیاز به الگوریتم‌های خوشه‌بندی گرافی وجود دارد که بتوانند به صورت هم‌زمان ساختار شبکه و ویژگی‌های گره‌ها را در فرآیند خوشه‌بندی در نظر بگیرند. به این دسته از روش‌ها، خوشه‌بندی گراف‌های دارای ویژگی^۱ گفته می‌شود [۱۰]. از سوی دیگر، با توجه به این‌که یک پروتئین می‌تواند در چندین فرآیند زیستی مختلف مشارکت داشته باشد، بسیاری از مجموعه‌های پروتئینی دارای هم‌پوشانی هستند؛ بنابراین الگوریتم مورد نظر باید قادر به شناسایی مجموعه‌های پروتئینی هم‌پوشان نیز باشد.

اگرچه بسیاری از پژوهش‌های پیشین تنها با تکیه بر ساختار شبکه‌های PPI و بدون در نظر گرفتن ویژگی‌های شناخته‌شده پروتئین‌ها، الگوریتم‌های کلاسیکی را برای شناسایی ماژول‌های عملکردی ارائه کرده‌اند، اما با پیشرفت‌های اخیر در حوزه هوش مصنوعی و الگوریتم‌های یادگیری ماشین، به‌ویژه یادگیری عمیق، ضرورت بررسی توانمندی شبکه‌های عصبی، به‌خصوص شبکه‌های عصبی گرافی، در این زمینه بیش از پیش احساس می‌شود. هرچند در برخی مطالعات، ترکیبی از روش‌های کلاسیک و شبکه‌های عصبی مورد استفاده قرار گرفته است، اما تاکنون یک رویکرد کاملاً مبتنی بر یادگیری عمیق و به صورت یکپارچه^۲ به طور جدی مورد توجه قرار نگرفته است.

در مجموع، هدف این پژوهش ارائه یک الگوریتم گراف‌محور برای کشف مجموعه‌های پروتئینی و ماژول‌های عملکردی هم‌پوشان در شبکه‌های PPI، با بهره‌گیری هم‌زمان از ساختار شبکه و داده‌های

¹ Attributed graph clustering

² End-to-end

تکمیلی استخراج شده از پایگاه‌های داده زیستی، در قالب چارچوب خوشه‌بندی گراف‌های دارای گره‌های ویژگی‌دار است. در ادامه، اهمیت موضوع، پرسش‌های اصلی پژوهش و روش پیشنهادی به صورت خلاصه مورد بررسی قرار خواهند گرفت.

۳.۱ اهمیت و ضرورت انجام پژوهش

در حال حاضر زیست‌شناسان توجه خود را از مطالعه ساختار و عملکرد انفرادی پروتئین‌ها به سمت مطالعه ساختاری و عملکردی مجموعه‌های پروتئینی تغییر داده‌اند و مولکول‌های پروتئین‌ها را درون یک شبکه زیستی کلی بررسی می‌کنند [۱۱]. دلیل این موضوع این است که بررسی یک پروتئین باید در ارتباط با سایر پروتئین‌ها و مجموعه‌های پروتئینی که به آن‌ها تعلق دارد، انجام شود. از سوی دیگر، جهش‌های موجود در دی‌ان‌ای می‌توانند نحوه برهم‌کنش پروتئین‌ها را در یک مجموعه پروتئینی تغییر دهند و به این ترتیب باعث تغییر در عملکرد و رفتار آن مجموعه شوند. این تغییرات نقش مهمی در بررسی فرآیند توسعه داروها و همچنین شناخت علل بروز بیماری‌ها دارند [۱۲، ۱۳]. به عنوان مثال پژوهش‌های [۱۱، ۱۴]، نشان داده‌اند که برخی از بیماری‌های ژنتیکی به وسیله پروتئین‌های با برهم‌کنش‌های عملکردی مشابه به وجود می‌آیند. همچنین مجموعه‌های پروتئینی به واسطه ارتباطشان با مسیرهای زیستی^۱، برای فهم بهتر نحوه توزیع، جذب، متابولیسم و دفع دارو ضروری هستند. از این روی شناسایی مجموعه‌های پروتئینی برای کشف و توسعه داروها اهمیت زیادی دارند. با وجود اهمیت کشف مجموعه‌های پروتئینی، چالش اصلی زمان‌بر و هزینه‌بر بودن فرآیند کشف آن‌ها در آزمایشگاه است که سبب توجه بیشتر محققان به روش‌های محاسباتی جهت کشف ماژول‌های عملکردی شده است [۱۵].

¹ Pathway

۴.۱ پرسش‌های پژوهش

بعد از تکمیل قسمت روش

۵.۱ روش پژوهش

بعدا تکمیل شود

۶.۱ جمع‌بندی

در این فصل به تعریف مسئله پژوهش و اهمیت آن پرداختیم. همچنین به صورت کلی روش پیشنهادی و سوالات پیش‌روی پژوهش را مورد بررسی قرار دادیم.

در ادامه، در فصل دوم به مبانی پایه مورد نیاز جهت فهم بهتر پژوهش اشاره خواهد شد. فصل سوم به معرفی پژوهش‌های پیشین که در زمینه شناسایی ماژول‌های عملکردی و مجموعه پروتئینی برپایه روش‌های محاسباتی هستند اختصاص دارد. فصل چهارم مربوط به روش شناسی و پژوهش است که در ابتدا داده‌های استفاده شده در این پژوهش سپس روش پیشنهادی خود را مطرح می‌کنیم. در فصل پنجم، به بررسی پیاده‌سازی روش پیشنهادی و نتایج حاصل بر اساس معیارهای ارزیابی می‌پردازیم. همچنین عملکرد روش پیشنهادی را با دیگر روش‌های پیشین مقایسه کرده و برتری روش خود را شرح می‌دهیم. در نهایت، در فصل آخر، پژوهش حاضر را جمع‌بندی می‌کنیم و ایده‌هایی را برای ادامه‌ی مسیر این پژوهش مطرح می‌کنیم.

فصل ۲

مفاهیم بنیادی

۱.۲ مقدمه

در این فصل، مفاهیم بنیادی و پیش‌نیازهایی که برای درک بهتر پژوهش حاضر ضروری هستند معرفی می‌شوند. از آنجا که این پایان‌نامه در تقاطع علوم زیستی و روش‌های محاسباتی قرار دارد، آشنایی با مفاهیم هر دو حوزه برای دنبال کردن مطالب فصل‌های بعدی اهمیت ویژه‌ای دارد. بر همین اساس، در بخش نخست این فصل، مفاهیم زیستی مرتبط با موضوع پژوهش مورد بررسی قرار می‌گیرند. سپس در بخش دوم، به معرفی مفاهیم محاسباتی مورد استفاده پرداخته می‌شود و تمرکز اصلی بر مباحث مرتبط با شبکه‌های عصبی گرافی و چارچوب‌های یادگیری مبتنی بر گراف خواهد بود. در نهایت، در بخش پایانی این فصل، معیارهای ارزیابی به کاررفته در این پژوهش برای سنجش کیفیت شناسایی ماژول‌های عملکردی معرفی شده و به‌طور خلاصه تشریح می‌شوند تا زمینه لازم برای تحلیل نتایج در فصل‌های بعدی فراهم شود.

۲.۲ مفاهیم زیستی

در این پژوهش، یک روش محاسباتی به منظور شناسایی مجموعه‌های پروتئینی ارائه می‌شود. در این راستا، با برخی مفاهیم زیستی مرتبط مواجه می‌شویم که در این بخش، توضیحات مختصر و روشنی از هر یک با هدف تسهیل درک موضوع پژوهش ارائه شده است.

۱.۲.۲ پروتئین

پروتئین‌ها از مهم‌ترین درشت مولکول‌های زیستی در سلول‌های زنده به‌شمار می‌آیند که از تکرار واحدهای اسید آمینه ساخته شده‌اند و نقش‌های حیاتی و متنوعی را در ساختار و عملکرد سلول ایفا می‌کنند. این مولکول‌ها به‌طور ویژه به‌عنوان کارگزاران مولکولی سلول شناخته می‌شوند، زیرا در فرآیندهایی مانند کاتالیز واکنش‌های شیمیایی (از طریق آنزیم‌ها)، انتقال مولکول‌ها، پیام‌رسانی درون‌سلولی، ایجاد حرکت و حفظ یکپارچگی ساختار سلولی نقش اساسی دارند. توالی اسیدهای آمینه هر پروتئین توسط ژن مربوطه تعیین شده و طی فرآیند ترجمه از روی آر‌ان‌ای (ریبونوکلئیک اسید)^۱ پیام‌رسان سنتز می‌شود. پس از سنتز، پروتئین‌ها از طریق فرآیند تاخوردگی^۲ به ساختارهای سه‌بعدی مشخصی دست می‌یابند که برای عملکرد زیستی آن‌ها ضروری است. ویژگی‌های عملکردی هر پروتئین به ترتیب خاص اسیدهای آمینه و برهم‌کنش‌های فضایی میان آن‌ها وابسته است. گستردگی و تنوع عملکردهای پروتئین‌ها به گونه‌ای است که تقریباً تمامی فرآیندهای زیستی سلول، به‌صورت مستقیم یا غیرمستقیم، تحت تأثیر یا کنترل آن‌ها قرار دارند [۱۶].

^۱ mRNA: messenger ribonucleic acid

^۲ Folding

۲.۲.۲ ماژول‌های عملکردی

فعالیت‌های زیستی در سلول و به‌طور کلی در بدن، معمولاً حاصل عملکرد یک پروتئین منفرد نیستند، بلکه نتیجه‌ی همکاری هماهنگ مجموعه‌ای از پروتئین‌ها می‌باشند که به‌صورت سازمان‌یافته با یکدیگر در ارتباط هستند. این پروتئین‌ها از طریق تعاملات مختلف، به‌ویژه تعاملات فیزیکی، در انجام یک یا چند وظیفه‌ی زیستی مشخص مشارکت می‌کنند [۱۵].



شکل ۱.۲: ساختار شماتیک یک گراف تعامل پروتئین-پروتئین، که نواحی رنگی ماژول‌های عملکردی را نشان می‌دهند.

به چنین مجموعه‌ای از پروتئین‌ها که به‌صورت هماهنگ برای انجام یک عملکرد زیستی مشترک عمل می‌کنند، مجموعه‌ی پروتئینی یا ماژول عملکردی^۱ گفته می‌شود. هر ماژول عملکردی معمولاً بیانگر یک فرآیند زیستی، مسیر مولکولی یا سازوکار تنظیمی خاص در سلول است و اجزای آن، از نظر عملکردی به یکدیگر وابسته‌اند [۱۷].

تعامل فیزیکی میان پروتئین‌ها که تحت عنوان تعامل پروتئین-پروتئین شناخته می‌شود، نقش محوری

¹ Functional Module

در شکل‌گیری و پایداری این ماژول‌های عملکردی ایفا می‌کند. این تعاملات امکان انتقال سیگنال، تنظیم فعالیت‌های آنزیمی و هماهنگی زمانی و مکانی پروتئین‌ها را فراهم می‌سازند و از این رو، برای درک صحیح بسیاری از فعالیت‌های زیستی ضروری هستند [۱۸].

از جمله فرآیندهای زیستی مهمی که مبتنی بر ماژول‌های عملکردی هستند می‌توان به رونوشت دی‌ان‌ای، رونوشت آر‌ان‌ای پیام‌رسان و تنظیم چرخه‌ی سلولی اشاره کرد. در هر یک از این فرآیندها، گروه مشخصی از پروتئین‌ها به صورت شبکه‌ای از تعاملات عمل می‌کنند و اختلال در هر یک از اجزای این شبکه می‌تواند منجر به بروز نقص عملکردی در کل فرآیند شود.

در سال‌های اخیر، پیشرفت در شناسایی و تحلیل ماژول‌های عملکردی، به یکی از موضوعات مهم در زیست‌شناسی سامانه‌ای و زیست‌انفورماتیک تبدیل شده است. شناسایی دقیق این ماژول‌ها کاربردهای گسترده‌ای از جمله پیش‌بینی عملکرد پروتئین‌های ناشناخته [۱۹]، درک سازوکارهای مولکولی بیماری‌ها [۲۰] و کشف اهداف دارویی جدید [۲۱] دارد. از این رو، مطالعه و مدل‌سازی ماژول‌های عملکردی نقش کلیدی در توسعه روش‌های نوین تشخیصی و درمانی ایفا می‌کند.

۳.۲.۲ بیان ژن

بیان ژن^۱ فرآیندی است که طی آن اطلاعات نهفته در توالی دی‌ان‌ای به محصولات عملکردی، عمدتاً آر‌ان‌ای و پروتئین، تبدیل می‌شود. این فرآیند شامل مراحل متعددی از جمله رونوشت دی‌ان‌ای به آر‌ان‌ای و در بسیاری از موارد ترجمه آر‌ان‌ای به پروتئین است و نقش اساسی در تعیین ساختار، عملکرد و رفتار سلول ایفا می‌کند. سطح بیان هر ژن نشان‌دهنده‌ی میزان فعالیت آن ژن در یک شرایط زیستی خاص بوده و به‌طور دقیق تحت تأثیر سازوکارهای تنظیمی مختلفی مانند عوامل رونویسی، تغییرات اپی‌ژنتیکی و سیگنال‌های درون‌سلولی و برون‌سلولی قرار دارد. تفاوت در الگوهای بیان ژن میان سلول‌ها، بافت‌ها یا شرایط فیزیولوژیک و پاتولوژیک مختلف، عامل اصلی تنوع عملکردی سلول‌ها محسوب می‌شود. از

^۱ Gene expression

این رو، تحلیل داده‌های بیان ژن ابزار مهمی برای درک فرآیندهای زیستی، شناسایی مسیرهای مولکولی مختل شده در بیماری‌ها و استخراج نشانگرهای زیستی به‌شمار می‌رود [۲۲].

۴.۲.۲ پایگاه داده هستی‌شناسی ژن

پایگاه داده هستی‌شناسی ژن یک بانک داده و سیستم طبقه‌بندی است که با هدف ایجاد یک زبان استاندارد برای توصیف ژن‌ها و محصولات ژنی (که پروتئین‌ها نیز جزو آنها هستند) ایجاد شده است. این پروژه اطلاعات ساختاریافته و قابل پردازش از فرآیندهای زیستی، عملکرد مولکولی و مؤلفه‌ی سلولی ژن‌ها فراهم می‌کند. داده‌های پروژه GO به صورت گسترده‌ای در تحقیقات مربوط به علوم زیستی مورد استفاده قرار می‌گیرد و همینطور همواره اطلاعات آن از نظر کمیت و کیفیت در حال تغییر است [۲۳].

Accession GO:0016597
 Name amino acid binding
 Ontology molecular_function
 Synonyms None
 Alternate IDs None
 Definition Interacting selectively and non-covalently with an amino acid, organic acids containing one or more amino substituents. Source: GOC:ai
 Comment None
 History See term [history for GO:0016597](#) at QuickGO
 Subset goslim_metagenomics
 goslim_pir
 Related [Link](#) to all **genes and gene products** annotated to amino acid binding.
[Link](#) to all direct and indirect **annotations** to amino acid binding.
[Link](#) to all direct and indirect **annotations download** (limited to first 10,000) for amino acid binding.

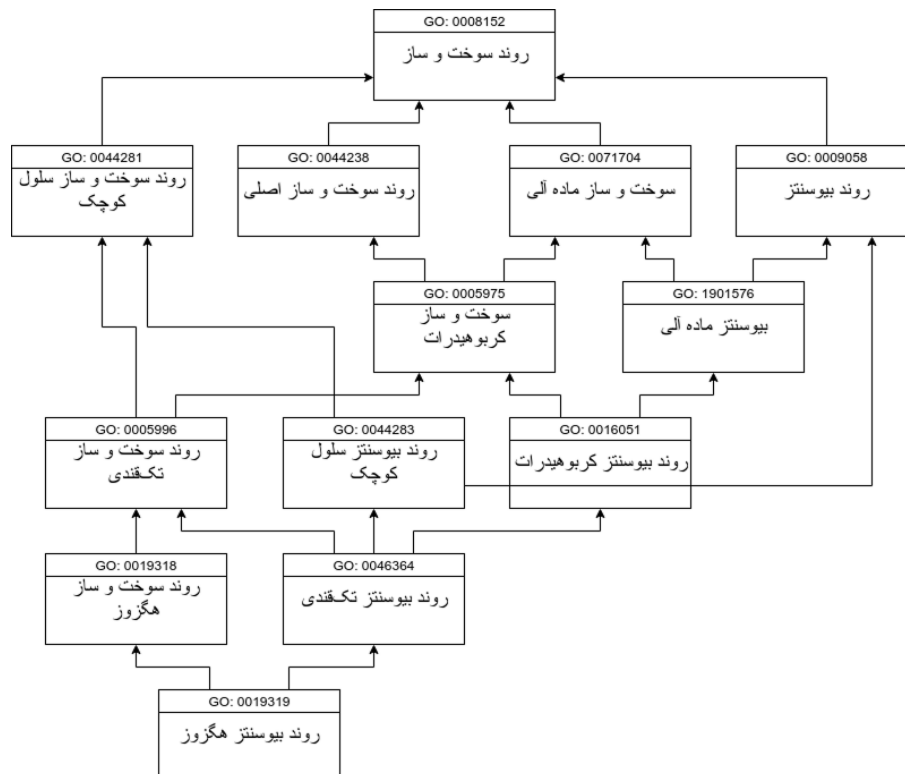
شکل ۲.۲: ویژگی‌های عبارت GO:0016597 [۲۴]

هر عبارت GO شامل موارد زیر می‌شود [۲۵]:

- یک نام که برای انسان قابل فهم باشد.
- یک شناساگر مختص آن عبارت که با پیشوند GO آغاز می‌شود.
- یک تعریف مختصر از مفاهیمی که توسط این عبارت GO نمایش داده می‌شود.
- ارتباط آن با سایر عبارات GO ؛ که در گراف GO هر عبارت (به جز عبارات ریشه‌ای) فرزند

یک عبارت GO دیگر است.

اطلاعات موجود در بانک داده GO به صورت ساختمان داده گرافی ذخیره شده‌اند. هر عبارت داری یک یا چند فرزند است که در نتیجه ساختار گراف GO، یک گراف جهت‌دار بدون دور^۱ است.



شکل ۳.۲: ساختار هستی‌شناسی ژن [۲۴]

گراف GO شامل چهار نوع یال `is_a`، `part_of`، `regulates` و `has_part` است که هر یک به‌ترتیب بیانگر رابطه «نوعی از»، «جزئی از»، «نقش تنظیم‌کنندگی» و «دارا بودن جزء» میان مفاهیم مختلف در این هستی‌شناسی می‌باشند [۲۶]. این سیستم شامل سه زیرگراف جهت‌دار بدون دور اصلی است که هر یک از آنها جنبه خاصی از عملکرد زیستی را توصیف می‌کنند:

- **فرآیند زیستی^۲** : این بخش به فرآیندهای زیستی اشاره دارد که ژن و یا پروتئین خاصی در آن

^۱ Directed acyclic graph

^۲ Biological process

نقش دارد.

- عملکرد مولکولی^۱ : این بخش عملکرد دقیق مولکولی ژن یا پروتئین را توصیف می‌کند.
- مؤلفه سلولی^۲ : این بخش به مکانی که ژن یا پروتئین در آن قرار دارد اشاره می‌کند. از ویژگی‌های دیگر این بانک داده نمایش اطلاعات به صورت سازماندهی شده و سلسله مراتبی است که شامل شبکه‌های بدون دور می‌شود و ویژگی‌ها به این صورت مرتب شده‌اند [۸].

۵.۲.۲ شبکه‌های PPI و ویژگی‌های آنها

یک شبکه PPI معمولاً به صورت یک گراف بدون جهت $G = (V, E)$ نشان داده می‌شود که V و E به ترتیب نمایانگر پروتئین‌ها و تعاملات بین آنها می‌باشند. وزن‌های روی یال‌ها را می‌توان برای توصیف ویژگی‌های شبکه PPI، مانند ویژگی‌های توپولوژیکی یا عملکردی استفاده کرد. شبکه‌های PPI سه ویژگی توپولوژیکی زیر را دارند:

- توزیع بدون مقیاس^۳ : $P(k)$ مفهوم توزیع درجه یعنی احتمال اینکه یک گره در یک شبکه دقیقاً k پیوند داشته باشد را نشان می‌دهد. یک شبکه PPI دارای توزیع درجه توانی $P(k) \sim k^{-\lambda}$ می‌باشد [۲۷]. این ویژگی به این معنی است که پروتئین‌های تعامل‌دار در شبکه‌های PPI به طور یکنواخت توزیع نمی‌شوند، بیشتر پروتئین‌ها تنها در چند تعامل شرکت می‌کنند در حالی که مجموعه کوچکی از پروتئین‌ها در ده‌ها تعامل (تشکیل گره‌هاب^۴) شرکت می‌کنند.
- ویژگی جهان کوچک^۵ : پروتئین‌های یک شبکه PPI دارای میانگین طول مسیر کم و ضرایب خوشه‌ای بالا هستند [۲۸] که سیگنال‌های هر گره در شبکه PPI را قادر می‌سازد تا از طریق چند

¹ Molecular function

² Cellular component

³ Scale-free distribution

⁴ Hub

⁵ Small-world property

جهش به سرعت به هر گره دیگری برسند. در نتیجه شبکه‌های PPI هم زمان انتقال سیگنال و هم زمان پاسخ کوتاهی خواهند داشت.

● شبکه با ماژول‌های عملکردی^۱: شبکه PPI یک شبکه ماژولار و سلسله مراتبی می‌باشد. یک ماژول عملکردی در یک شبکه PPI یک مجموعه با بیشترین تعداد پروتئین که عملکرد یکسانی دارند، می‌باشد. بارزترین مشخصه ماژول عملکردی، ارتباط بین ساختار توپولوژیکی شبکه PPI و عملکرد پروتئین‌های آن است که مبنای بسیاری از روش‌های تشخیص ماژول عملکردی است [۲۹] [۳۰].

۳.۲ مفاهیم محاسباتی

در این بخش، مفاهیم محاسباتی مورد استفاده در این پایان‌نامه معرفی می‌شوند. از آن‌جا که روش پیشنهادی این پژوهش بر پایه تحلیل شبکه‌ها و یادگیری عمیق استوار است، آشنایی با مبانی نظری مرتبط با گراف‌ها، الگوریتم‌های یادگیری عمیق و شبکه‌های عصبی گرافی برای درک بهتر مراحل روش ارائه شده ضروری است. بدین منظور، در ادامه مروری اجمالی بر مفاهیم و تعاریف اصلی ارائه می‌شود تا چارچوب محاسباتی پژوهش به صورت منسجم و شفاف تبیین گردد.

۱.۳.۲ یادگیری ماشین

یادگیری ماشین^۲ یکی از شاخه‌های هوش مصنوعی^۳ است که به رایانه‌ها امکان می‌دهد بدون تعریف صریح قوانین، الگوها و روابط موجود در داده‌ها را شناسایی کرده و بر اساس آن‌ها به پیش‌بینی یا تصمیم‌گیری بپردازند. در این رویکرد، مدل‌ها با استفاده از داده‌های آموزشی، دانش لازم را استخراج

¹ Functional modular network

² Machine Learning

³ Artificial Intelligence

کرده و قادر خواهند بود این دانش را به داده‌های جدید تعمیم دهند. امروزه یادگیری ماشین در بسیاری از حوزه‌های پژوهشی و صنعتی مورد استفاده قرار می‌گیرد. به‌طور کلی، روش‌های یادگیری ماشین به سه دسته‌ی اصلی یادگیری نظارت‌شده^۱، یادگیری بدون نظارت^۲ و یادگیری تقویتی^۳ تقسیم می‌شوند که هر یک متناسب با نوع داده‌ها و هدف مسئله به‌کار گرفته می‌شوند [۳۱، ۳۲].

۲.۳.۲ یادگیری عمیق

یادگیری عمیق را می‌توان یکی از زیرشاخه‌های یادگیری ماشین دانست که بر پایه‌ی مدل‌هایی با چندین لایه‌ی پردازشی بنا شده است. این مدل‌ها معمولاً از لایه‌های متصل به‌هم یا لایه‌های پیچیده‌تر تشکیل می‌شوند و به‌دلیل ساختار چندلایه‌ی خود، دارای تعداد پارامترهای قابل یادگیری بیشتری نسبت به الگوریتم‌های کلاسیک یادگیری ماشین هستند. همین ویژگی موجب می‌شود که یادگیری عمیق توانایی بالاتری در مدل‌سازی روابط پیچیده میان داده‌ها داشته باشد.

تفاوت اصلی یادگیری عمیق با روش‌های سنتی یادگیری ماشین در نحوه‌ی استخراج ویژگی‌ها است. در الگوریتم‌های کلاسیک، ویژگی‌ها معمولاً به‌صورت دستی و با دخالت متخصص تعیین می‌شوند، در حالی که روش‌های یادگیری عمیق قادرند با توجه به داده‌های موجود و ماهیت مسئله، ویژگی‌های مناسب را به‌صورت خودکار استخراج کنند. این قابلیت باعث شده است که یادگیری عمیق در مسائل پیچیده و داده‌محور عملکرد بهتری از خود نشان دهد [۳۳].

^۱ Supervised Learning

^۲ Unsupervised Learning

^۳ Reinforcement Learning

۳.۳.۲ یادگیری نظارت شده

یادگیری نظارت شده یکی از رایج ترین انواع یادگیری ماشین است که در آن الگوریتم با استفاده از داده های ورودی به همراه برچسب^۱ متناظر با هر نمونه آموزش می بیند. هدف از این فرآیند، ساخت مدلی است که بتواند رابطه میان داده ها و برچسب ها را آموخته و برای داده های جدید و دیده نشده عملکرد مناسبی داشته باشد. در این نوع یادگیری، داده ها معمولاً به دو مجموعه ی آموزشی^۲ و آزمایشی^۳ تقسیم می شوند؛ بدین صورت که مدل با استفاده از داده های آموزشی ساخته شده و سپس با داده های آزمایشی مورد ارزیابی قرار می گیرد. به بیان ساده، وجود برچسب در کنار هر داده، ویژگی اصلی یادگیری نظارت شده محسوب می شود [۳۴].

۴.۳.۲ یادگیری بدون نظارت

یادگیری بدون نظارت نوعی از یادگیری ماشین است که در آن الگوریتم ها با داده های بدون برچسب سروکار دارند و هدف اصلی آن ها کشف الگوها و ساختارهای پنهان موجود در داده ها است. برخلاف یادگیری نظارت شده، در این روش اطلاعاتی از خروجی مطلوب در اختیار مدل قرار نمی گیرد و الگوریتم تلاش می کند بر اساس شباهت ها و ویژگی های ذاتی داده ها، آن ها را سازمان دهی کند. از شناخته شده ترین الگوریتم های یادگیری بدون نظارت می توان به روش های خوشه بندی اشاره کرد که با هدف قرار دادن داده های مشابه در خوشه های یکسان به کار می روند [۳۵].

۵.۳.۲ گراف

یک گراف از مجموعه ای غیر خالی از اشیا به نام رأس تشکیل شده، که آن را با V نشان می دهیم، و مجموعه ای شامل یال ها، که رأس ها را به هم وصل می کنند و با E نمایش می دهیم. یک چنین

¹ Label

² Train dataset

³ Test dataset

گرافی را با $G = (V, E)$ نشان می‌دهیم. اگر یال e دو رأس v_1 و v_2 را به هم وصل کند می‌نویسیم $e = \{v_1, v_2\}$ [۳۶]. تعریف ارائه شده، تعریف گراف ساده است. اما انواع مختلفی از گراف موجود می‌باشد که در ادامه به بررسی دو نوع از آن‌ها (گراف جهت‌دار و گراف وزن‌دار) می‌پردازیم:

• **گراف جهت‌دار:** گراف $G(V, E)$ زمانی جهت‌دار است که مجموعه E ، از جفت‌های (u, v) تشکیل شده باشد و ترتیب این دو تایی‌ها نشان‌دهنده جهت یال همانند $u, v \in V$ باشد. به این صورت برای هر یال جهت نیز در نظر گرفته می‌شود که به گراف حاصل، گراف جهت‌دار می‌گوییم.

• **گراف وزن‌دار:** گراف $G(V, E, W)$ که $W \in R^{|E|}$ یک مقدار عددی به هر یک از یال‌ها اختصاص می‌دهد که میزان وزن آن یال است.

۶.۳.۲ شبکه‌های عصبی گرافی

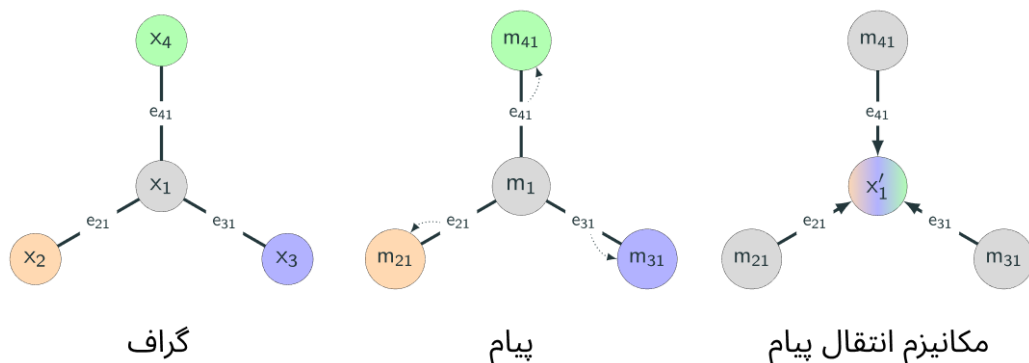
شبکه‌های عصبی گرافی^۱ اولین بار در سال ۲۰۰۵ پیشنهاد شدند [۳۷]. شبکه‌های عصبی گرافی، دسته‌ای از شبکه‌های عصبی هستند که برای مدیریت داده‌های سازمان‌دهی شده در ساختارهای گراف طراحی شده‌اند. شبکه‌های عصبی گرافی بر پایه سازوکار انتقال پیام^۲ هستند.

در ابتدا یک گراف با ماتریس ویژگی گره‌ها $X \in R^{|V| \times d}$ به عنوان ورودی در نظر گرفته می‌شود که $|v|$ تعداد گره‌های گراف و d بعد ویژگی‌های گراف می‌باشد. در شبکه‌های عصبی گرافی از این ویژگی‌ها در کنار ساختار گراف برای تولید تعبیه‌های هر گره استفاده می‌شود. در هر تکرار، هر گره اطلاعاتی را از گره‌های همسایگی خود جمع‌آوری می‌کند که این عمل را به صورت کلی جمع‌آوری^۳ می‌نامند. در مرحله بعد شبکه باید اطلاعات جمع‌آوری شده را با اطلاعات موجود گره ادغام کند و تعبیه جدیدی

¹ Graph neural networks - GNNs

² Message passing

³ Aggregate



شکل ۴.۲: شماتیک سازوکار انتقال پیام در شبکه‌های عصبی گرافی [۳۸]

از گره مورد نظر ارائه دهد. به صورت کلی این مرحله از انتقال پیام را نیز بروزرسانی^۱ می‌نامند. به طور خلاصه در یکبار انتقال پیام مراحل زیر طی می‌شوند:

$$h_u^{(k+1)} = UPDATE^{(k)}(h_u^{(k)}, AGGREGATE(\{h_v^{(k)}, \forall v \in N(u)\})) \quad (۱.۲)$$

در فرمول ۱.۲ نمادهای AGGREGATE و UPDATE، دو تابع دلخواه مشتق‌پذیر (به عنوان مثال یک شبکه عصبی) هستند و $N(u)$ نشانگر مجموعه همسایگان گره u است. همچنین $h_u^{(k)}$ نشان‌دهنده تعبیه گره u در مرحله k ام است.

با افزایش این مراحل، تعبیه هر گره داده‌های بیشتری از گره‌های دورتر از خود در گراف خواهد داشت. پس از اولین تکرار ($k = 1$)، هر تعبیه گره اطلاعات مربوط به همسایگی تک گامی خود را حفظ می‌کند، که ممکن است در گراف از طریق مسیری به طول ۱ قابل دسترسی باشد [۳۹]. بعد از دومین تکرار ($k = 2$)، تعبیه هر گره شامل اطلاعاتی از همسایگی با دو گام است؛ به طور کلی، پس از k

¹ Update

مرحله، تعبیه هر گره می‌تواند شامل داده‌هایی از گره‌هایی با فاصله $k - hop$ از خود باشد. براساس سازوکار انتقال پیام در شبکه‌های عصبی گرافی، این شبکه‌ها در تولید تعبیه‌هایی که هم اطلاعات مربوط به ساختار گراف و هم ویژگی‌های گره‌ها را حفظ کنند بسیار موفق بوده‌اند. به همین دلیل در بسیاری از مسائل مورد استفاده قرار گرفته‌اند. بنابر توابع استفاده شده به عنوان تابع جمع آوری و بروزرسانی، شبکه‌های عصبی گرافی به انواع مختلفی همانند شبکه‌های عصبی گرافی پیچشی، شبکه‌های عصبی گرافی توجه محور و دیگر دسته‌ها تقسیم بندی می‌شوند [۴۰].

۷.۳.۲ شبکه‌های عصبی گرافی پیچشی

پژوهش کیف و ولینگ [۴۱] با هدف ارائه مدلی ساده، مقیاس پذیر و قابل اجرا برای یادگیری روی گراف‌های بزرگ، شبکه‌های عصبی گرافی پیچشی را معرفی کرد. روش‌های طیفی پیشین مبتنی بر تجزیه‌ی لاپلاسین گراف بوده و نیازمند محاسبه‌ی مقادیر ویژه و بردارهای ویژه بودند که این امر هزینه‌ی محاسباتی بالایی داشته و استفاده از آن‌ها را در گراف‌های بزرگ محدود می‌کرد. شبکه گرافی پیشنهاد شده از معادله ۲.۲ برای انتقال پیام استفاده می‌کند.

$$H^{(k+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^k W^k) \quad (2.2)$$

در اینجا، $\tilde{A} = A + I_N$ ماتریس مجاورت گراف بدون جهت G با در نظر گرفتن یال‌های خودی^۱ است. I_N ماتریس همانی بوده و $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ عناصر قطری ماتریس درجه متناظر را تشکیل می‌دهد. همچنین، $W^{(k)}$ یک ماتریس وزن قابل آموزش مخصوص لایه k ام است. تابع $\sigma(\cdot)$ نشان‌دهنده یک تابع فعال‌سازی است که برای مثال می‌تواند تابع ReLU به صورت $\text{ReLU}(\cdot) = \max(\cdot, 0)$ باشد. ماتریس $H^{(k)} \in \mathbb{R}^{N \times D}$ نمایش دهنده فعال‌سازی‌ها در لایه k ام بوده و $H^{(0)} = X$ به عنوان ورودی اولیه شبکه در نظر گرفته می‌شود. با توجه به تعریف ارائه شده در بخش شبکه‌های عصبی گرافی و سازوکار

¹ Self-connections

انتقال پیام، توابع AGGREGATE و UPDATE، به ترتیب از معادله‌های $S = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(k)}$ و $\sigma(SW^{(k)})$ پیروی می‌کنند.

با وجود سادگی و کارایی بالا، شبکه‌های عصبی گرافی پیچشی دارای محدودیت‌هایی نیز هستند. از جمله این محدودیت‌ها می‌توان به پدیده‌ی هموارسازی بیش‌ازحد^۱ در صورت افزایش تعداد لایه‌ها اشاره کرد که در آن بردار تعبیه گره‌ها به مرور شباهت زیادی به یکدیگر پیدا می‌کنند و توان تفکیک مدل کاهش می‌یابد. علاوه براین، در این شبکه‌ها، تمامی گره‌های همسایه به یک اندازه در تشکیل تعبیه جدید نقش دارند و این در حالی است که در بسیاری از گراف‌ها میزان اهمیت تمامی گره‌ها یکسان نیست. این محدودیت‌ها انگیزه‌ای برای توسعه‌ی مدل‌های پیشرفته‌تر شبکه‌های عصبی گرافی در پژوهش‌های بعدی بوده است.

۸.۳.۲ شبکه‌های عصبی گرافی توجه محور

ولیکویک و همکاران [۴۲]، با هدف رفع محدودیت‌های شبکه‌های عصبی گرافی پیچشی در تخصیص وزن یکسان به تمامی همسایه‌ها، مدل شبکه‌های عصبی گرافی با سازوکار توجه^۲ را معرفی کردند. این مدل امکان یادگیری وزن‌های متفاوت برای هر یال همسایگی را فراهم می‌کند و بدین ترتیب اهمیت نسبی گره‌های همسایه در تشکیل بردار تعبیه هر گره به صورت داده‌محور تعیین می‌شود. این ویژگی باعث می‌شود GAT توانایی مدل‌سازی پیچیدگی‌های ساختاری گراف‌هایی با ارتباطات غیرهمگن را داشته باشد، در حالی که GCN تمامی همسایه‌ها را با وزن یکسان در نظر می‌گیرد.

در GAT، انتقال پیام با استفاده از سازوکار توجه خود-توجهی^۳ در معادله ۳.۲ تعریف می‌شود.

¹ Over-smoothing

² Graph Attention Network — GAT

³ Self-attention

$$h_i^{(k+1)} = \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} W h_j^{(k)} \right) \quad (3.2)$$

که در آن، $h_i^{(k+1)}$ نمایش به‌روزشده گره i ، $h_j^{(k)}$ ویژگی‌های گره همسایه j ، W ماتریس وزن قابل‌آموزش و $\sigma(\cdot)$ تابع فعال‌سازی است. ضریب توجه α_{ij} اهمیت گره j را نسبت به گره i مشخص می‌کند و با استفاده از یک شبکه کوچک خطی و تابع LeakyReLU مطابق معادله ۴.۲ محاسبه می‌شود.

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(a^T [W h_i^{(k)} \| W h_j^{(k)}]))}{\sum_{n \in \mathcal{N}(i)} \exp(\text{LeakyReLU}(a^T [W h_i^{(k)} \| W h_n^{(k)}]))} \quad (4.2)$$

که در آن a بردار وزن قابل‌آموزش برای سازوکار توجه و $\|$ عملگر ادغام ویژگی‌های گره‌ها است. به این ترتیب، تابع AGGREGATE در GAT به صورت جمع‌وزنی گره‌های همسایه با ضرایب توجه بوده و تابع UPDATE شامل اعمال ماتریس وزن و تابع فعال‌سازی روی مقدار جمع‌وزنی شده است.

یکی از نوآوری‌های مهم GAT استفاده از توجه چندسر^۱ است. در این روش، K سازوکار توجه مستقل بر روی همان لایه اعمال می‌شود و نتایج هر سر یا با هم ادغام می‌شوند (الحاق^۲ یا میانگین‌گیری) تا نمایی غنی‌تر و پایدارتر از ویژگی‌های گره‌ها تولید شود که در معادله ۵.۲ نمایش داده شده است.

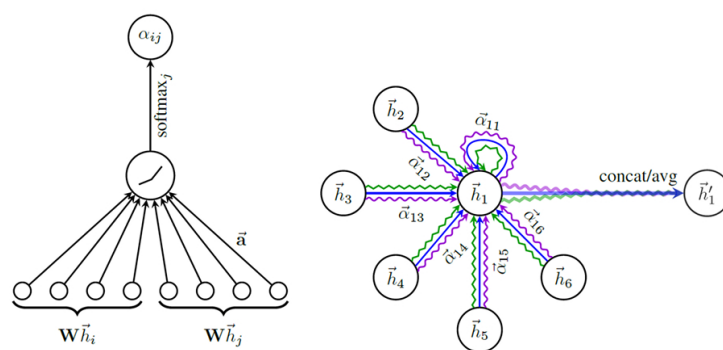
$$h_i^{(k+1)} = \parallel_{m=1}^M \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(m)} W^{(m)} h_j^{(k)} \right) \quad (5.2)$$

استفاده از سازوکار توجه چند سر، علاوه بر افزایش ظرفیت مدل، باعث کاهش حساسیت شبکه به نویز

¹ Multi-head attention

² Concat

و نوسانات محلی گراف می‌شود و کمک می‌کند تا مدل بتواند اطلاعات مفیدی از همسایگان مختلف در سطوح گوناگون استخراج کند.



شکل ۵.۲: سازوکار توجه در شبکه‌های عصبی گرافی [۴۲]

با وجود مزایای قابل توجه، GAT نیز محدودیت‌هایی دارد. پیچیدگی محاسباتی آن نسبت به GCN بالاتر است، به ویژه در گراف‌های بزرگ با درجه بالا. همچنین، انتخاب تعداد سرهای توجه و تنظیمات مربوط به آنها می‌تواند تأثیر زیادی بر عملکرد شبکه داشته باشد و نیازمند تنظیمات دقیق است. با این حال، GAT به خوبی امکان مدل‌سازی اهمیت متفاوت همسایگان، کاهش اثرات هموارسازی بیش‌ازحد و استخراج ویژگی‌های غیرهمگن را فراهم می‌کند و به همین دلیل در بسیاری از مسائل یادگیری روی گراف، نتایج بهتری نسبت به GCN ارائه می‌دهد.

۹.۳.۲ شبکه‌های عصبی گرافی با انتقال دانش بین لایه‌ها

شبکه‌های عصبی گرافی با انتقال دانش بین لایه‌ها^۱ برای نخستین بار در پژوهش ژو و همکاران [۴۳] معرفی شدند. این پژوهش به یکی از چالش‌های اساسی در شبکه‌های عصبی گرافی اشاره می‌کند که ناشی از استفاده از تعبیه‌های مبتنی بر فاصله‌های ثابت همسایگی برای تمام گره‌ها است. در چنین رویکردی، تمامی گره‌ها صرف‌نظر از موقعیت ساختاری خود در گراف، با تعداد یکسانی از لایه‌ها پردازش می‌شوند؛ در حالی که ساختار زیرگرافی گره‌ها می‌تواند به‌طور قابل توجهی با یکدیگر متفاوت

^۱ Jumping Knowledge Graph Neural Networks - JKNNets

باشد. نویسندگان این پژوهش بر این باورند که به منظور استخراج تعبیه مناسب‌تر برای هر گره، لازم است مرتبه‌ی تعبیه آن متناسب با ویژگی‌های ساختاری همان گره انتخاب شود. به بیان دیگر، برخی گره‌ها برای دستیابی به تعبیه معنادار به اطلاعات محلی نیاز دارند، در حالی که برای برخی دیگر، بهره‌گیری از همسایگی‌های دورتر ضروری است.

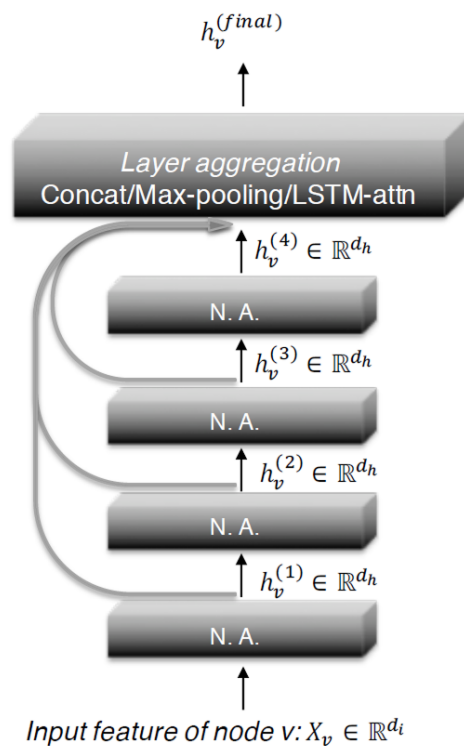
برای نمونه، در پژوهش کیف و همکاران [۴۱] که منجر به معرفی شبکه‌های عصبی گرافی پیچشی شد، نتایج تجربی نشان می‌دهد که استفاده از دو لایه بهترین عملکرد را به همراه دارد. این در حالی است که از منظر تئوری، افزایش تعداد لایه‌ها باید امکان تجمع اطلاعات گسترده‌تر و یادگیری تعبیه‌های غنی‌تر را فراهم کند. این ناسازگاری بیانگر وجود محدودیت‌هایی در تعمیق شبکه‌های عصبی گرافی است. در حوزه‌ی بینایی ماشین، مشکل تعمیق شبکه‌ها با بهره‌گیری از سازوکار اتصال باقیمانده^۱ تا حد زیادی برطرف شده است [۴۴]. با این حال، در شبکه‌های عصبی گرافی، حتی با وجود استفاده از اتصال‌های باقیمانده، مسئله‌ی هموار شدن بیش‌ازحد تعبیه گره‌ها و محدودیت شعاع همسایگی همچنان باقی می‌ماند. در همین راستا، ژو و همکاران با هدف رفع مشکل شعاع ثابت همسایگی و افزایش انعطاف‌پذیری در استخراج بردار تعبیه گره‌ها، سازوکار انتقال دانش بین لایه‌ها را پیشنهاد کردند. این سازوکار امکان ترکیب تطبیقی اطلاعات حاصل از لایه‌های مختلف شبکه را برای هر گره فراهم می‌کند و بدین ترتیب، هر گره می‌تواند از سطح مناسبی از اطلاعات محلی یا سراسری بهره‌مند شود.

سازوکار انتقال دانش بین لایه‌ها مستقل از نوع معماری بوده و قابلیت اعمال بر روی انواع شبکه‌های عصبی گرافی را دارد. با در نظر گرفتن تعبیه‌های استخراج‌شده از لایه‌های مختلف شبکه برای گره v به صورت $h_v^{(1)}, h_v^{(2)}, \dots, h_v^{(k)}$ ، این سازوکار اقدام به ترکیب این تعبیه‌ها با استفاده از روش‌های مختلفی نظیر بیشینه‌گیری^۲، الحاق^۳، یا به کارگیری شبکه‌های حافظه‌ی کوتاه‌مدت بلندمدت مبتنی بر

¹ Residual Connection

² Max Pooling

³ Concatenation



شکل ۶.۲: شماتیک پرس دانش در شبکه‌های عصبی گرافی [۴۳]

سازوکار توجه^۱ می‌کند. بدین ترتیب، بردار تعبیه نهایی هر گره به صورت تطبیقی و متناسب با ساختار آن در گراف شکل می‌گیرد.

۱۰.۳.۲ تعبیه گره‌ها به روش Node2Vec

روش Node2Vec یک الگوریتم مقیاس پذیر^۲ نیمه نظارتی^۳ برای یادگیری ویژگی‌ها از روی گراف است. این الگوریتم به طور مستقیم از الگوریتم یادگیری تعبیه کلمات Word2Vec [۴۵] که در زمینه پردازش زبان طبیعی^۴ استفاده می‌شود، ایده گرفته است. در این روش هدف تابع بهینه‌سازی، بیشینه کردن احتمال مشاهده گره‌های همسایه یک گره به شرط مشاهده خود آن گره است. هدف نهایی

^۱ LSTM-attention

^۲ Scalable

^۳ Semi-Supervised

^۴ Natural Language Processing - NLP

این الگوریتم یادگیری یک بردار تعبیه d بعدی برای هر گره است [۴۶]. این روش در ابتدا اقدام به جایگشت تصادفی بر روی گراف به کمک الگوریتم‌های نمونه برداری اول سطح^۱ و اول عمق^۲ می‌کند. انتخاب روش مناسب نمونه برداری از اهمیت بالایی برخوردار است. در نمونه برداری اول سطح، هدف اصلی استخراج تعبیه‌های مشابه برای گره‌هایی است که از قوانین ساختاری یکسانی پیروی می‌کنند؛ در حالی که نمونه برداری اول عمق بر ایجاد تعبیه‌های مشابه برای گره‌هایی تمرکز دارد که به صورت چگال به یکدیگر متصل هستند. در عمل، بهترین راهکار استفاده از یک روش ترکیبی است؛ به گونه‌ای که بخشی از توالی‌ها با استفاده از نمونه برداری اول عمق و بخش دیگری با استفاده از نمونه برداری اول سطح تولید شوند. سپس یک شبکه عصبی آموزش داده می‌شود تا با استفاده از مشاهده مسیرهای پیشین، گره بعدی را پیش‌بینی کند (مشابه فرآیند آموزش در روش Word2Vec). بدین منظور، تابع هزینه^۳ مطابق با معادله ۶.۲ تعریف می‌شود.

$$\max_f \sum_{u \in V} \log Pr(N_S(u) | f(u)) \quad (۶.۲)$$

۱۱.۳.۲ خوشه‌بندی گراف با گره‌های ویژگی‌دار

با فرض گراف $G = (V, E, F)$ که در آن V مجموعه گره‌ها، E مجموعه یال‌ها است و F ماتریس ویژگی‌های گره‌ها می‌باشد، یک خوشه بندی از گراف G را می‌توان با C نشان داد که مجموعه‌ای از زیر مجموعه‌های V است، به صورتی که $C_i \in C$; $C_i \subset V$. هدف از خوشه بندی این است که خوشه‌هایی که هم از نظر ساختاری و هم از نظر ویژگی‌های گره‌ها بهم بیشترین شباهت را دارند، پیدا کنیم. همچنین خوشه‌های ایجاد شده باید از نظر ارتباط یال‌های داخل خوشه چگال و در ارتباط یال‌ها با دیگر خوشه‌ها تنک باشند. نکته مهم دیگر در این قسمت وجود و یا عدم وجود همپوشانی در

¹ Breadth-first sampling - BFS

² Depth-first sampling - DFS

³ Loss function

بین خوشه‌ها می‌باشد که به خوشه بندی بدون همپوشانی، افزایش‌بندی^۱ نیز می‌گویند. به عبارت دیگر در افزایش‌بندی شرط:

$$\forall i, j; i \neq j; C_i \in C \text{ and } C_j \in C; C_i \cap C_j \subseteq \phi$$

باید حتما رعایت شود این در حالیست که در خوشه‌بندی با همپوشانی چنین شرطی الزامی نیست.

۱۲.۳.۲ دسته‌بندی و روش‌های کلی خوشه‌بندی گراف

روش‌های خوشه‌بندی گراف را می‌توان از دیدگاه‌های مختلفی تقسیم‌بندی کرد. این تقسیم‌بندی‌ها بر اساس معیارها و ویژگی‌های خاصی صورت می‌گیرند که به نحوه برخورد با داده‌های گرافی، نوع اطلاعات استفاده شده، و تکنیک‌های به کار گرفته شده بستگی دارد. در این پژوهش از آنجایی که نوع گراف ورودی مشخص است و قصد خوشه‌بندی گراف‌های PPI با گره‌های دارای ویژگی را داریم، روش‌های خوشه بندی را بر اساس روش مورد استفاده تقسیم‌بندی می‌کنیم:

- روش‌های طیفی^۲: از مقادیر ویژه^۳ ماتریس لاپلاسی یا مجاورت برای یافتن خوشه‌ها استفاده می‌کنند.
- روش‌های فاکتورگیری ماتریسی^۴: از روش‌های تجزیه ماتریسی مانند تجزیه نامنفی ماتریس^۵ یا تجزیه مقدار تکین^۶ برای ایجاد بردار تعبیه و خوشه‌بندی استفاده می‌کنند.
- روش‌های سلسله‌مراتبی^۷: گراف را به صورت سلسله مراتبی خوشه‌بندی می‌کنند که به دو روش تقسیمی و تجمعی دسته‌بندی می‌شوند.

¹ Partitioning

² Spectral clustering

³ Eigenvalues

⁴ Matrix factorization

⁵ Non-negative matrix factorization

⁶ Singular value factorization

⁷ Hierarchical clustering

- روش‌های مبتنی بر تعبیه^۱ : ابتدا گره‌ها به فضای برداری کم‌بعد نگاشت می‌شوند و سپس خوشه‌بندی روی این فضای برداری انجام می‌شود و تمرکز اصلی در این روش‌ها یافتن بردار تعبیه مناسب برای خوشه‌بندی گراف است. الگوریتم‌های CNN، GCN، Deep Walk و Node2Vec جزء این دسته محسوب می‌شوند.

- روش‌های بدون تعبیه^۲ : مستقیماً از ساختار گراف برای خوشه‌بندی استفاده می‌شود بدون اینکه گره‌ها به فضای برداری منتقل شوند. به عنوان مثال می‌توان روش‌های مبتنی بر graph-cut و یا روش Louvain را نام برد.

۴.۲ معیارهای ارزیابی

در این قسمت به بررسی معیارهای ارزیابی عملکرد الگوریتم‌های شناسایی مجموعه‌های پروتئینی می‌پردازیم. در بین معیارهای موجود، معیارهای دقت^۳، بازیابی^۴، صحت^۵، امتیاز F، بیشترین استفاده را در بین پژوهش‌ها داشته‌اند که ما نیز به منظور تحلیل و مقایسه عملکرد روش خود از آنها استفاده می‌کنیم. در ابتدا برای شروع به معیار شباهت همسایگی که برای محاسبه تمامی معیارهای مذکور مورد نیاز است، می‌پردازیم:

۱.۴.۲ شباهت همسایگی

در تعریف شباهت همسایگی^۶، با در نظر گرفتن P به عنوان مجموعه‌ای از مجموعه‌های پروتئینی شناسایی شده توسط الگوریتم، عملکرد الگوریتم به وسیله تعداد مجموعه‌های پروتئینی مشترک بین

¹ Embedding-based methods

² Non-embedding methods

³ Precision

⁴ Recall

⁵ Accuracy

⁶ Neighborhood affinity

P و مجموعه‌ای از مجموعه پروتئین‌های مرجع^۱ B بدست می‌آید. برای مشخص کردن اینکه آیا یک مجموعه پروتئین شناسایی شده $p \in P$ با یک مجموعه پروتئین مرجع $b \in B$ یکسان هستند یا خیر ما اقدام به محاسبه معیار شباهت همسایگی به صورت مقابل می‌کنیم:

$$NA(p, b) = \frac{|V_p \cap V_b|^2}{|V_p| \times |V_b|} \quad (۷.۲)$$

که V_p مجموعه پروتئین‌های حاضر در ترکیب p و به طور مشابه V_b مجموعه پروتئین‌های حاضر در b هستند. برای تفسیر شباهت همسایگی یک آستانه^۲ از قبل تعیین شده (معمولاً ۰/۲۵) در نظر گرفته می‌شود که شباهت همسایگی‌های بالاتر از آستانه به معنی یکسانی دو مجموعه است. همچنین تعداد مجموعه‌های شناسایی شده‌ای که حداقل با یک مجموعه مرجع یکسان در نظر گرفته می‌شوند را با N_{cp} و تعداد مجموعه‌های مرجعی که حداقل با یکی از مجموعه‌های شناسایی شده الگوریتمی یکسان در نظر گرفته می‌شوند را با N_{cb} نمایش می‌دهیم [۴].

$$N_{cp} = \{p | p \in P, \exists b \in B, NA(p, b) \geq \omega\} \quad (۸.۲)$$

$$N_{cb} = \{b | b \in B, \exists p \in P, NA(p, b) \geq \omega\} \quad (۹.۲)$$

¹ Reference protein complex

² Threshold

۲.۴.۲ دقت

دقت یک معیار ارزیابی مجموعه پروتئینی‌های شناسایی شده است که نشان می‌دهد چند مورد از مجموعه‌های پیش‌بینی شده الگوریتم به درستی انتخاب شده‌اند.

$$Precision = \frac{N_{cp}}{|P|} \quad (۱۰.۲)$$

۳.۴.۲ بازیابی

بازیابی دیگر معیار مورد توجه است که نشان می‌دهد چند مورد از مجموعه پروتئینی‌های مرجع توسط الگوریتم پیش‌بینی شده‌اند. به دیگر عبارت میزان پوشش الگوریتم از مجموعه پروتئینی‌های مرجع را اندازه‌گیری می‌کند.

$$Recall = \frac{N_{cb}}{|B|} \quad (۱۱.۲)$$

۴.۴.۲ امتیاز F

معیار امتیاز F میانگین همساز^۱ بین دو معیار دقت و بازیابی می‌باشد که به صورت مقابل محاسبه می‌شود:

$$F - score = \frac{۲ \times Precision \times Recall}{Precision + Recall} \quad (۱۲.۲)$$

^۱ Harmonic mean

۵.۴.۲ صحت

معیار صحت به کمک دو معیار دیگر حساسیت خوشه‌بندی^۱ و ارزش پیش‌بینی مثبت خوشه‌بندی^۲ محاسبه می‌شود. با در نظر گرفتن $T_{i,j}$ به عنوان تعداد پروتئین‌هایی که هم در مجموعه پروتئینی i ام و هم در مجموعه پروتئینی پیش‌بینی j ام یافت می‌شوند و همچنین N به عنوان تعداد پروتئین‌های مجموعه پروتئینی مرجع i ، می‌توانیم PPV و Sn را با معادله ۱۳.۲ تعریف کنیم.

$$PPV = \frac{\sum_{j=1}^{|P|} \max_{i=1}^{|B|} |T_{ij}|}{\sum_{j=1}^{|P|} \sum_{i=1}^{|B|} T_{ij}} \quad (13.2)$$

$$Sn = \frac{\sum_{i=1}^{|B|} \max_{j=1}^{|P|} |T_{ij}|}{\sum_{i=1}^{|B|} N_i}$$

از نظر مفهومی، معیار PPV نشان‌دهنده‌ی نسبت مجموع بیشینه پروتئین‌های تطبیق‌یافته هر مجموعه پیش‌بینی شده با مجموعه‌های پروتئینی مرجع، به تعداد کل پروتئین‌های تطبیق‌یافته در مجموعه‌های پروتئینی پیش‌بینی شده است. از سوی دیگر، معیار Sn بیان‌کننده‌ی نسبت مجموع بیشینه پروتئین‌های تطبیق‌یافته هر مجموعه پروتئینی مرجع با مجموعه‌های پروتئینی پیش‌بینی شده، به تعداد کل پروتئین‌های موجود در مجموعه‌های پروتئینی مرجع است. در نهایت به کمک این دو معیار می‌توان معیار صحت را به صورت مقابل محاسبه نمود:

$$Acc = \sqrt{Sn.PPV}$$

¹ Clustering-wise sensitivity (Sn)

² Clustering-wise positive predictive value (PPV)

فصل ۳

بررسی منابع

در این قسمت به بررسی پژوهش‌های پیشینی که به منظور پیدا کردن مجموعه‌های پروتئینی در شبکه‌های PPI انجام شده‌اند، می‌پردازیم. همان‌طور که در بخش‌های پیشین بررسی شد، تمرکز این پژوهش بر روی دید گرافی به شبکه‌های PPI و ادغام اطلاعات زیست‌شناسی پروتئین‌ها به منظور تشخیص دقیق‌تر مجموعه‌های پروتئینی است. از آنجایی که پیدا کردن مجموعه‌های پروتئینی در شبکه‌های PPI معادل خوشه‌بندی این شبکه‌ها می‌باشد، ما ابتدا چند نمونه از پژوهش‌های مرتبط با خوشه‌بندی گراف‌های دارای گره ویژگی که بیشترین ارتباط را با هدف پژوهش ما دارند را بررسی می‌کنیم.

۱.۳ خوشه‌بندی گراف با گره‌های ویژگی‌دار

پژوهش وحید جان‌نثاری و همکارانش [۴۷]، یک الگوریتم بر پایه تجزیه نامنفی ماتریسی^۲ به منظور خوشه‌بندی گراف‌های ویژگی‌دار معرفی می‌کند. روش آن‌ها ابتدا اطلاعات ساختاری که توسط ماتریس

² Non-negative matrix factorization

همسایگی^۱ نشان داده می‌شود را به کمک تجزیه نامنفی متقارن ماتریس^۲ و اطلاعات ویژگی‌های گره‌ها را به کمک تجزیه نامنفی بازتابی ماتریس^۳ به یک فضای کم بعد مختص خوشه‌بندی (هم بعد با تعداد خوشه‌ها) به صورت جداگانه انتقال می‌دهد که درجه عضویت هر گره به هر خوشه را نمایش می‌دهد. همین‌طور به منظور حفظ ثبات در خوشه‌بندی در هر دو فضا اقدام به نزدیک کردن این دو ماتریس به کمک تابع هدف می‌کند که به صورت مقابل تعریف شده است:

$$J_{of} = \min \|A - VV^T\|_F^2 + \alpha \|VV^T - UU^T\|_F^2 + \|F - UU^T F\|_F^2 \quad (۱.۳)$$

$$s.t. \ V \geq 0, \ U \geq 0, \ V^T V = I, \ U^T U = I.$$

که در تابع هدف، A ماتریس همسایگی، $V \in R^{n \times k}$ ماتریس حاصل از تجزیه نامنفی متقارن ماتریس A است. همین‌طور با در نظر گرفتن M (ماتریس شباهت^۴ گره‌ها براساس ماتریس ویژگی‌ها) به صورت $M = UU^T; U \in R^{n \times k}$ و عبارت سوم در بهینه‌سازی که به صورت مقابل بیان شده است: $\|F - MF\|$ در واقع اقدام به استفاده از ویژگی خودبیانگری^۵ داده‌ها کرده‌اند، که در نتیجه روش بیان شده را می‌توان یک روش ترکیبی از خوشه‌بندی زیر فضا^۶ و تجزیه نامنفی ماتریس در نظر گرفت.

در پژوهشی دیگر توسط کانگ و همکارانش [۴۸]، یک روش بر پایه شبکه‌های پیچشی گرافی^۷ و خوشه‌بندی طیفی ارائه شده است. ایده اصلی در این روش بر پایه پردازش سیگنالی گراف است که در آن یک فیلتر پایین گذر^۸ را به منظور نزدیک کردن و ادغام ویژگی‌های گره‌ها و ساختار گراف به ماتریس

¹ Adjacency matrix

² Symmetric non-negative matrix factorization

³ Projective non-negative matrix factorization

⁴ Similarity matrix

⁵ Self-expression

⁶ Subspace clustering

⁷ Graph convolutional networks

⁸ Low-pass filter

ویژگی‌ها اعمال می‌کنند. در نتیجه یک تعبیه جدید بر این اساس را برای گره‌ها بدست می‌آورند:

$$\bar{X} = (I - \frac{1}{2}L)^k X \quad (2.3)$$

همچنین در معادله بالا k یک هاپر پارامتر است که میزان مرتبه مجاورت تعبیه به دست آمده را مشخص می‌کند به عبارت دیگر مقادیر کوچک‌تر k دید محلی‌تری به ساختار گراف دارند و بالعکس. L ماتریس لاپلاسی نرمال شده^۱ است که به صورت $L = I - A$ تعریف می‌شود. در مرحله بعد برای اعمال خوشه‌بندی طیفی، نیاز به محاسبه ماتریس شباهت بین گره‌ها است که به صورت مقابل عمل کرده‌اند.

$$\min_S ||\bar{X}^T - \bar{X}^T S||_F^2 + ||S - f(A)||_F^2 \quad (3.3)$$

که در اینجا ماتریس شباهت S از بهینه سازی تابع هدف بالا بدست می‌آید و سپس با یک انتقال به یک ماتریس متقارن نامنفی تبدیل شده و در نهایت نیز خوشه‌بندی طیفی روی آن اعمال می‌شود. یکی از مشکلات این روش انتخاب مناسب هاپر پارامتر K است که به طور مستقیم بر خروجی الگوریتم تأثیر می‌گذارد که توسط پژوهش دیگری که توسط ژانگ و همکارانش [۴۹] انجام شده است، دو استراتژی AGC و IAGC برای پیدا کردن مقدار مناسب k ارائه شده است.

بهومیک و همکاران [۵۰] روشی مبتنی بر بیشینه‌سازی ماژولاریتی با عنوان DGCluster ارائه داده‌اند که در آن از شبکه‌های عصبی گرافی برای خوشه‌بندی گراف‌های ویژگی‌دار استفاده می‌شود. در این چارچوب، ابتدا یک شبکه عصبی گرافی پیچشی دولایه با تابع فعال‌ساز SELU به منظور ایجاد بردارهای تعبیه گره‌ها به کار گرفته می‌شود. سپس با اعمال مجموعه‌ای از تبدیلات غیرخطی و نرمال‌سازی، تعبیه‌های حاصل به فضای مختصات مثبت محدود می‌گردند تا محاسبه شباهت بین گره‌ها تسهیل شود. در ادامه، با استفاده از شباهت کسینوسی میان بردارهای تعبیه گره‌ها، نسخه تغییر یافته‌ای از

¹ Normalized Laplacian matrix

ماژولاریتی تعریف شده و تابع هزینه متناظر با آن به صورت منفی ماژولاریتی معادله بندی می شود.

$$L_1 = -\hat{Q} = \frac{1}{m} Tr(BXX^T) \quad (4.3)$$

که در آن هر درایه ماتریس B به صورت $B_{ij} = (A_{ij} - \frac{d_i d_j}{m})$ و X ماتریس تعبیه هر گره حاصل از شبکه عصبی گرافی بعد می باشد. پس از آموزش شبکه عصبی با هدف بیشینه سازی ماژولاریتی، فرآیند خوشه بندی نهایی بر روی بردارهای تعبیه آموخته شده و با استفاده از الگوریتم خوشه بندی سلسله مراتبی BIRCH انجام می گیرد که نیازی به تعیین تعداد خوشه ها از قبل ندارد [51].

در پژوهش هی و همکاران [52]، یک چارچوب یکپارچه با عنوان SSAGCN¹ برای خوشه بندی گراف ها معرفی شده است. در این روش، ابتدا گراف ورودی ویژگی دار به صورت $G(V, E, A, X)$ در نظر گرفته می شود. سپس گراف کمکی دیگری به شکل $G(V, E_a, A_a, X)$ ساخته می شود که تفاوت آن با گراف اولیه در مجموعه یال ها و ماتریس مجاورت است. این گراف کمکی بر اساس محاسبه شباهت بین ویژگی های گره ها با استفاده از معیار شباهت کسینوسی تشکیل شده و در آن هر گره به k گره مشابه خود متصل می شود. در ادامه، هر دو گراف به دو شبکه عصبی گرافی پیچشی با وزن های به اشتراک گذاشته شده وارد شده و بردارهای تعبیه حاصل از آن ها با بهره گیری از یک مکانیزم توجه برای هر گره ترکیب می شوند. تابع هزینه پیشنهادی به صورت ترکیب خطی و مبتنی بر بیشینه سازی ماژولاریتی، بازسازی ماتریس مجاورت و بازسازی ماتریس ویژگی ها تعریف شده است که از آن با عنوان مکانیزم رمزگشای دوگانه² یاد می شود. در نهایت، با استفاده از تعبیه های نهایی به دست آمده، میزان تعلق هر گره به خوشه های مختلف محاسبه شده و هر گره به خوشه ای اختصاص می یابد که بیشترین میزان تعلق را داشته باشد.

یکی از مشکلات روش های بر پایه تعبیه این است که دو فرآیند تعبیه داده ها و خوشه بندی از یکدیگر

¹ Self-Supervised Adaptive Graph Convolutional Network

² Dual Decoder

مستقل‌اند در نتیجه نمی‌توان اطمینان داشت که تعبیه‌های ایجاد شده برای وظیفه موردنظر (در اینجا خوشه‌بندی) مناسب هستند و همچنین نمی‌توان الگوریتم تعبیه را بر اساس خطای خوشه‌بندی به طور مناسب به روزرسانی نمود. از این روی، وانگ و همکاران [۵۳] یک روش خوشه‌بندی یکپارچه توجه محور بر پایه شبکه عصبی گراف ارائه داده‌اند که مرحله تعبیه و خوشه‌بندی را با هم ترکیب می‌کند. در این پژوهش از یک شبکه گرافی توجه محور^۱ به عنوان کدگذار استفاده شده است. ضرایب توجه کدگذار^۲ با استفاده از یک ماتریس مجاورت با مرتبه بالا همانند پژوهش قبلی محاسبه می‌شوند. قسمت کدگشا^۳ نیز از ضرب داخلی بردارهای تعبیه کدگذار به منظور بازسازی ماتریس مجاورت گراف استفاده می‌کند که برای خروجی این قسمت تابع هزینه بازسازی در نظر گرفته شده است. نوآوری این مقاله در معرفی مفهوم تعبیه خود بهینه‌ساز است که در آن به طور مکرر نقاط مربوط به هر خوشه براساس مقدار اطمینان تعلق به خوشه به‌روزرسانی می‌شود و به طور همزمان تعبیه‌ها را نیز به وسیله آن اصلاح می‌کند. از جمله پژوهش‌های شاخص در زمینه خوشه‌بندی گراف‌های ویژگی‌دار^۴ می‌توان به کار شچور^۵ و گانمن^۶ اشاره نمود [۵۴] که به‌عنوان مبنای اصلی این پایان‌نامه نیز مورد استفاده قرار گرفته است. در این پژوهش، یک الگوریتم یکپارچه برای شناسایی خوشه‌های هم‌پوشان در گراف‌ها به نام NOCD ارائه شده است که ایده اصلی آن بر تلفیق توانمندی‌های شبکه‌های عصبی گرافی با یک مدل مولد احتمالی برنولی-پواسون^۷ استوار است.

در چارچوب این روش، گراف ورودی به‌صورت $G(A, X)$ در نظر گرفته می‌شود که در آن ماتریس مجاورت به صورت $A \in \{0, 1\}^{n \times n}$ و $X \in \mathbb{R}^{n \times d}$ ماتریس ویژگی گره‌ها است. هدف، آموزش یک شبکه عصبی گرافی با پارامترهای θ ، موسوم به GNN_θ ، به‌منظور استخراج ماتریس عضویت خوشه‌ای

¹ Graph attentional

² Decoder

³ Encoder

⁴ Attributed Graphs

⁵ Shchur

⁶ Günnemann

⁷ Probabilistic Bernoulli-Poisson Generative Model

F می باشد:

$$F = GNN_{\theta}(A, X) \quad (5.3)$$

در این رابطه، $F \in \mathbb{R}^{n \times C}_{\geq}$ ماتریس انتساب گره‌ها به خوشه‌ها بوده و هر مؤلفه F_{uc} بیانگر میزان یا شدت تعلق گره u به خوشه c است. شبکه عصبی گرافی به کاررفته در این پژوهش، یک شبکه گرافی پیچشی دولایه با تابع فعال‌ساز ReLU است که مطابق معادله ۶.۳ تعریف می‌شود. در این ساختار، با افزودن یال‌های خوداتصال و نرمال‌سازی ماتریس مجاورت، اطلاعات ساختاری و ویژگی‌های گره‌ها به‌صورت همزمان در فرآیند یادگیری مورد استفاده قرار می‌گیرند.

$$\begin{aligned} \tilde{A} &= A + I_N \\ \tilde{D}_{ii} &= \sum_j \tilde{A}_{ij} \\ \hat{A} &= \tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} \\ F &= ReLU(\hat{A} ReLU(\hat{A} X W^{(1)}) W^{(2)}) \end{aligned} \quad (6.3)$$

پس از استخراج ماتریس F ، یک مدل مولد برنولی-پواسون به‌عنوان رمزگشا به کار گرفته می‌شود که هدف آن بازسازی ماتریس مجاورت گراف است. در این مدل، وجود یال میان دو گره u و v به‌صورت یک متغیر تصادفی برنولی با پارامتر $1 - e^{-F_u F_v^T}$ مدل‌سازی می‌شود:

$$A_{uv} \sim Bernoulli(1 - e^{-F_u F_v^T}) \quad (7.3)$$

بر این اساس، احتمال مشاهده ماتریس مجاورت A با شرط دانستن ماتریس عضویت F به صورت

ضرب احتمال‌ها بر روی یال‌های موجود و ناموجود محاسبه می‌شود:

$$P(A|F) = \prod_{A_{uv} \in E} (1 - e^{-F_u F_v^T}) \times \prod_{A_{uv} \notin E} e^{-F_u F_v^T} \quad (۸.۳)$$

در نهایت، تابع هزینه مدل به صورت $-\log p(A|F)$ تعریف شده و به شکل زیر قابل بیان است:

$$L(F) = - \sum_{A_{uv} \in E} \log(1 - \exp(-F_u F_v^T)) + \sum_{A_{uv} \notin E} F_u F_v^T \quad (۹.۳)$$

پژوهشگران در این مقاله به یک چالش اساسی در تابع هزینه فوق اشاره می‌کنند. از آن‌جا که گراف‌های واقعی عموماً تنک^۱ هستند، تعداد جفت‌گره‌هایی که بین آن‌ها یال وجود ندارد به مراتب بیشتر از یال‌های موجود است؛ در نتیجه، عبارت دوم تابع هزینه غالب شده و می‌تواند فرآیند یادگیری را تحت تأثیر قرار دهد. به منظور رفع این مشکل، نویسندگان با استفاده از امید ریاضی هر یک از عبارات تابع هزینه و فرض یک توزیع یکنواخت بر روی یال‌ها، نسخه اصلاح‌شده‌ای از تابع هزینه را ارائه می‌دهند که منجر به پایداری بیشتر و بهبود عملکرد مدل در گراف‌های تنک می‌شود.

$$L(F) = -E_{(U,V) \sim P_E} [\log(1 - \exp(-F_u F_v^T))] + E_{(u,v) \sim P_N} [F_u F_v^T] \quad (۱۰.۳)$$

که در این معادله عبارت P_E توزیع یکنواخت بر روی یال‌ها و P_N توزیع یکنواخت بر روی گره‌هایی است که یالی بین آن‌ها وجود ندارد. پس از آموزش شبکه عصبی، برای پیدا کردن خوشه‌ها با استفاده

^۱ Sparse

از ماتریس وابستگی F از آستانه φ استفاده می‌شود.

$$F_{uc} = \begin{cases} 1 & \text{if } F_{uc} > \varphi \\ 0 & \text{otherwise} \end{cases} \quad (11.3)$$

۲.۳ پیش‌بینی مجموعه‌های پروتئینی

در ادامه به بررسی روش‌های استفاده شده به منظور پیش‌بینی مجموعه‌های پروتئینی در شبکه‌های PPI می‌پردازیم و یک دسته‌بندی برای این روش‌ها ارائه می‌دهیم.

۱.۲.۳ روش‌های بر پایه شبکه

این روش‌ها تنها بر ساختار شبکه PPI تمرکز می‌کنند. که به دو زیر دسته تقسیم می‌شوند:

- روش‌های تقسیمی^۱: این دسته از روش‌ها، شبکه را به زیر شبکه‌ها تقسیم می‌کنند و این عمل را تا رسیدن به درجه دلخواه خوشه بندی تکرار می‌کنند. معروف ترین الگوریتم این دسته الگوریتم خوشه‌بندی مارکوف^۲ [۵۵] است که زیر شبکه‌ها را به کمک ولگشت تصادفی^۳ در شبکه پیدا می‌کند.

- روش‌های تجمعی^۴: با مجموعه کوچکی از پروتئین‌ها شروع کرده و با ترکیب آن‌ها اقدام به پیدا کردن مجموعه‌های پروتئینی نهایی می‌کند. الگوریتم CPNM [۵۶] یکی از الگوریتم‌های این دسته است که از تعبیه موتیف‌های^۵ شبکه به منظور پیدا کردن نقش پروتئین‌ها استفاده

¹ Divisive methods

² Markov clustering algorithm

³ Random walk

⁴ Agglomerative methods

⁵ Motif

می‌کند. سپس به منظور ایجاد بردار ویژگی پروتئین‌ها از آن‌ها استفاده می‌شود. در نهایت نیز از روش پیدا کردن همسایگان به منظور شناسایی مجموعه‌های پروتئینی استفاده می‌کند. یکی دیگر از الگوریتم‌های تجمعی معروف الگوریتم ClusterONE [۵۷] است. این الگوریتم ابتدا پروتئین‌های با درجه بالاتر را به عنوان پروتئین‌های هسته^۱ (پروتئین‌های آغازین) در نظر می‌گیرد. سپس زیرگروه‌هایی از گره‌ها با بیشترین انسجام برای گره‌های هسته انتخاب می‌شوند. در انتها نیز گره هسته از بین گره‌هایی که مربوط به یک ترکیب شناخته شده نیستند انتخاب می‌شوند و این مراحل تکرار می‌شوند تا همه پروتئین‌ها به یک ترکیب مرتبط شوند. الگوریتم دیگر، MCODE^۲ [۵۸] است که در سه مرحله انجام می‌شود. این الگوریتم ابتدا گره‌ها را وزن دهی می‌کند، سپس به شناسایی مجموعه‌ها می‌پردازد و در انتها نیز اقدام به اضافه / حذف کردن پروتئین‌ها به/از مجموعه‌های شناسایی شده با توجه به یک معیار اتصال می‌کند.

۲.۲.۳ روش‌های مبتنی بر آگاهی از زمینه‌های زیستی

اگرچه روش‌های بر پایه شبکه عملکرد خوبی دارند، اما عملکرد آنها می‌تواند با به کارگیری اطلاعات تکمیلی بهبود یابد. این اطلاعات می‌توانند از منابع گوناگونی مثل اطلاعات دامنه‌ای پروتئین‌ها، برچسب‌های ژن شناسی، نمایه بیان ژنی جمع آوری شوند. پژوهش آلن و همکارانش [۵۹]، الگوریتم PCIA را توسعه داده‌اند که از ترکیب اطلاعات GO در کنار ساختار شبکه استفاده می‌کند. پژوهش دیگر ژانگ و همکارانش [۶۰] رابطه‌ی بین شکل‌گیری مجموعه‌های پروتئینی و هم بیانی پروتئین‌ها را نشان داده است.

- روش‌های هسته-اتصال^۳: روش‌های هسته-اتصال بر پایه این ایده هستند که هر مجموعه پروتئینی از یک هسته تشکیل شده است که شامل پروتئین‌هایی با هم بیانی بالا می‌باشند.

¹ Seed

² Molecular complex detection

³ Core-attachment

الگوریتم COACH [۶۱] یکی از شناخته شده ترین الگوریتم های این دسته است که از دو مرحله شناسایی پروتئین های هسته ای و اضافه کردن پروتئین ها به پروتئین های هسته ای تشکیل شده است. تمرکز این الگوریتم بر ایجاد مجموعه های پروتئینی است که از نظر زیستی نیز با معنی باشند. الگوریتم CORE [۶۲] نیز از سه مرحله، پیش بینی پروتئین های هسته ای، حذف هسته های با اهمیت پایین (بر اساس یک معیار اتصال)، و محاسبه اهمیت مجموعه های شناسایی شده، تشکیل شده است. اخیرا نیز الگوریتم CO-DPC از این دسته بندی ارائه شده است که از نمایه بیان ژنی در کنار شبکه PPI استفاده می کند.

● الگوریتم های مبتنی بر اطلاعات عملکردی^۱: دسته دوم الگوریتم ها روش های مبتنی بر اطلاعات عملکردی هستند که از اطلاعات ناهمگون پروتئین ها به منظور شناسایی مجموعه های با معنی استفاده می کنند. یکی از الگوریتم های این دسته، الگوریتم PCP [۶۳] است که از اطلاعات ساختاری به منظور وزن دهی شبکه PPI استفاده می کند. سپس ابتدا اقدام به شناسایی کلیک های بیشینه^۲ در شبکه PPI کرده، در مرحله بعد چگالی بین خوشه ها را محاسبه می کند و در نهایت اقدام به ترکیب جزئی کلیک ها می کند. از دیگر پژوهش های شاخص در این بخش می توان به پژوهش برهمند و همکارانش [۶۴] اشاره کرد. این پژوهش با تأکید بر این نکته که شناسایی دقیق تر ماژول های عملکردی مستلزم بهره گیری همزمان از ساختار شبکه و ویژگی های زیستی پروتئین هاست، روشی با عنوان TADW-SC را پیشنهاد کرده اند. در این روش، علاوه بر ساختار شبکه PPI، از ویژگی های استخراج شده از GO نیز استفاده می شود. بدین منظور، در گام نخست با بهره گیری از روش TADW^۳ [۶۵]، بردارهای تعبیه ای برای هر گره محاسبه می شوند که به گونه ای طراحی شده اند تا همزمان اطلاعات ساختاری گراف و ویژگی های گره ها را حفظ کنند. در مرحله بعد، با استفاده از بردارهای تعبیه آموخته شده، یک ماتریس شباهت بین گره ها

¹ Functional-information-based

² Maximal clique

³ Text Associated Deep Walk

ایجاد می‌شود. سپس الگوریتم خوشه‌بندی طیفی بر روی این ماتریس شباهت اعمال شده و ماژول‌های عملکردی استخراج می‌گردند. نتایج تجربی نشان می‌دهد که روش TADW-SC با وجود ساختار ساده، در بسیاری از شبکه‌های PPI عملکرد قابل قبولی از خود نشان داده و می‌تواند به‌عنوان یک رویکرد کارا برای شناسایی ماژول‌های عملکردی مورد استفاده قرار گیرد. لازم به ذکر است که استراتژی‌های دیگری که در سایر پژوهش‌های مربوط به پردازش گراف‌ها و خوشه‌بندی آنها خوب عمل کرده‌اند نیز مورد توجه قرار گرفته‌اند که از جمله آنها می‌توان به روش‌های بر پایه تعبیه [۶۶] [۶۷] و تجزیه ماتریسی [۶۸] اشاره کرد که به منظور شناسایی مجموعه‌های پروتئینی نیز مورد استفاده قرار گرفته‌اند.

۳.۲.۳ روش‌های مبتنی بر شبکه‌های عصبی گرافی

در این میان، روش‌های مبتنی بر شبکه‌های عصبی گرافی برای شناسایی مجموعه‌های پروتئینی به‌طور محدود مورد مطالعه قرار گرفته‌اند و بررسی و توسعه این دسته از روش‌ها یکی از اهداف اصلی پژوهش حاضر به‌شمار می‌رود. با این حال، نتایج امیدوارکننده‌ای از به‌کارگیری شبکه‌های عصبی گرافی در این حوزه گزارش شده است؛ به‌گونه‌ای که روش پیشنهادی چن و همکاران [۶۹] در مقایسه با بسیاری از رویکردهای پیشین، بهترین عملکرد را در شناسایی ماژول‌های عملکردی شبکه‌های PPI نشان داده است. در پژوهش چن و همکاران، چارچوبی با عنوان AdaPPI را برای شناسایی مجموعه‌های پروتئینی در شبکه‌های برهم‌کنش PPI معرفی می‌کند. این چارچوب ترکیبی از یک شبکه عصبی گرافی پیچشی و یک راهبرد یادگیری تعبیه تطبیقی در سطح گره است که با هدف استخراج تعبیه‌هایی متناسب با ساختار محلی و سراسری شبکه طراحی شده است. در این روش، پس از استخراج تعبیه‌های گره‌ها، شناسایی ماژول‌های عملکردی با بهره‌گیری از تعبیه‌های آموخته‌شده و با استفاده از الگوریتم کلاسیک یافتن کلیک‌های هسته‌ای و گسترش آنها انجام می‌شود.

در AdaPPI، ویژگی‌های مربوط به GO هر پروتئین شامل فرآیندهای زیستی و عملکردهای مولکولی،

از زیرگراف‌های متناظر استخراج شده و به صورت بردارهای دودویی کدگذاری می‌شوند. این ویژگی‌ها به عنوان ورودی اولیه به شبکه عصبی گرافی پیچشی داده می‌شوند. در ادامه، خروجی هر لایه از شبکه عصبی گرافی به یک شبکه عصبی بازگشتی وارد می‌شود که وظیفه آن پیش‌بینی میزان هموارسازی مناسب برای تعبیه هر گره است. این شبکه بازگشتی^۱ به صورت همزمان با شبکه عصبی گرافی آموزش داده می‌شود.

ایده اصلی این سازوکار بر این فرض استوار است که گره‌های مختلف در شبکه PPI به سطوح متفاوتی از اطلاعات همسایگی نیاز دارند. به طور خاص، گره‌های با درجه بالا عمدتاً از اطلاعات همسایگان مرتبه پایین‌تر بهره می‌برند، در حالی که گره‌های با درجه پایین‌تر برای به‌روزرسانی تعبیه خود نیازمند دسترسی به اطلاعات همسایگان با مراتب بالاتر هستند. راهبرد یادگیری تعبیه تطبیقی در AdaPPI از بروز پدیده‌ی هموارسازی بیش از حد یا کمتر از حد تعبیه‌ها جلوگیری کرده و امکان استخراج تعبیه‌های گره‌ای متمایزتر را فراهم می‌سازد. پس از یادگیری تعبیه‌ها، با استفاده از ساختار گراف و تعبیه‌های حاصل، کلیک‌های هسته‌ای در شبکه شناسایی شده و سپس از طریق الگوریتم گسترش کلیک، ماژول‌های عملکردی نهایی استخراج می‌شوند. تابع هزینه مورد استفاده در این پژوهش به گونه‌ای طراحی شده است که فاصله بین تعبیه گره‌های متعلق به یک خوشه را کاهش داده و همزمان فاصله تعبیه گره‌های متعلق به خوشه‌های متفاوت را افزایش دهد. این تابع هزینه به صورت معادله ۱۲.۳ تعریف می‌شود:

$$\begin{aligned}
 L_{\text{intra}} &= \frac{1}{|C|} \sum_{c \in C} \frac{1}{|c|(|c| - 1)} \sum_{i, j \in c, i \neq j} \|\bar{X}_i - \bar{X}_j\|_2 \\
 L_{\text{inter}} &= \frac{1}{|C|} \sum_{c \in C} \frac{1}{|c|(n - |c|)} \sum_{i \in c, j \notin c} \|\bar{X}_i - \bar{X}_j\|_2 \\
 L &= L_{\text{intra}} + \beta \frac{1}{L_{\text{inter}}}
 \end{aligned} \tag{12.3}$$

با توجه به اینکه چارچوب پیشنهادی AdaPPI از ترکیب شبکه‌های عصبی و الگوریتم‌های کلاسیک

¹ Recurrent Neural Network - RNN

خوشه‌بندی بهره‌می‌برد و فرآیند شناسایی ماژول‌های عملکردی در دو مرحله مجزا انجام می‌شود، این روش در دسته الگوریتم‌های یکپارچه قرار نمی‌گیرد. با این حال، نتایج تجربی نشان می‌دهد که این روش در بسیاری از شبکه‌های PPI عملکرد بهتری نسبت به روش‌های پیشین داشته و به‌عنوان یکی از مؤثرترین الگوریتم‌ها برای شناسایی ماژول‌های عملکردی در این حوزه مطرح است.

فصل ۴

روش شناسی پژوهش

۱.۴ مقدمه

پژوهش حاضر به شناسایی ماژول‌های عملکردی پروتئینی در شبکه‌های PPI با بهره‌گیری از شبکه‌های عصبی گرافی می‌پردازد. همان‌گونه که در بخش بیان مسئله مطرح شد، مسئله شناسایی ماژول‌های عملکردی پروتئینی را می‌توان به صورت یک مسئله خوشه‌بندی در شبکه‌های PPI مدل‌سازی کرد. از این رو، در فصل پژوهش‌های پیشین، مروری بر مطالعات مرتبط با خوشه‌بندی گراف و روش‌های شناسایی ماژول‌های عملکردی ارائه شد.

بررسی مطالعات پیشین نشان می‌دهد که با وجود پیشرفت‌های قابل توجه در حوزه خوشه‌بندی شبکه‌های زیستی، تعداد پژوهش‌هایی که به طور مستقیم از روش‌های مبتنی بر شبکه‌های عصبی گرافی برای تحلیل شبکه‌های PPI استفاده کرده‌اند، همچنان محدود است. افزون بر این، تاکنون چارچوب یادگیری یکپارچه و منسجمی که به صورت خاص برای شناسایی ماژول‌های عملکردی در شبکه‌های PPI طراحی شده باشد، به طور کامل مورد توجه قرار نگرفته است. این پژوهش در پی آن است تا با بررسی دقیق‌تر

ظرفیت‌ها و قابلیت‌های شبکه‌های عصبی گرافی، گامی در جهت پر کردن این خلأ پژوهشی بردارد. از آن‌جا که مسئله شناسایی ماژول‌های عملکردی ذاتاً یک مسئله یادگیری بدون نظارت به شمار می‌رود، انتخاب تابع هزینه مناسب و تعیین معیارهای توقف مؤثر نقش بسزایی در تضمین همگرایی و کارایی فرآیند یادگیری ایفا می‌کنند. علاوه بر این، روش پیشنهادی باید قادر باشد به‌صورت هم‌زمان اطلاعات ساختاری شبکه و ویژگی‌های زیستی پروتئین‌ها را به‌طور مؤثر مدلسازی کند. در این راستا، در نظر گرفتن همپوشانی میان خوشه‌ها از اهمیت ویژه‌ای برخوردار است؛ چرا که در شبکه‌های زیستی واقعی، ماژول‌های عملکردی پروتئینی اغلب دارای همپوشانی قابل توجهی با یکدیگر هستند.

این فصل به دو بخش کلی تقسیم می‌شود. در بخش نخست، مجموعه داده‌های مورد استفاده معرفی می‌شوند که شامل شبکه‌های PPI مورد بررسی، پایگاه داده هستی‌شناسی ژن، مراحل پیش‌پردازش داده‌ها و فرآیند آماده‌سازی آن‌ها است. در بخش دوم، روش پیشنهادی این پژوهش به‌طور جامع ارائه شده و اجزای مختلف آن به‌صورت دقیق مورد تحلیل و بررسی قرار می‌گیرند.

۲.۴ مجموعه داده

در دهه‌های اخیر، شبکه‌های برهم‌کنش پروتئین-پروتئین (Protein-Protein Interaction یا PPI) به‌واسطه توسعه روش‌های آزمایشگاهی با توان عملیاتی بالا^۱ به‌طور قابل توجهی گسترش یافته‌اند. از جمله این روش‌ها می‌توان به سیستم‌های دوگانه هیبریدی^۲ [۷۰] و طیف‌سنجی جرمی^۳ [۷۱] اشاره کرد. افزون بر این، روش‌های مبتنی بر متن‌کاوی^۴ نیز به‌صورت گسترده برای استخراج تعاملات پروتئینی و ایجاد شبکه‌های PPI مورد استفاده قرار گرفته‌اند [۷۳، ۷۲، ۸].

از منظر مدلسازی گرافی، این منابع داده امکان نمایش شبکه‌های PPI را به‌صورت یک گراف بدون

^۱ High-throughput

^۲ Two-hybrid systems

^۳ Mass spectrometry

^۴ Text mining

جهت $G = (V, E)$ فراهم می‌کنند که در آن هر گره $v \in V$ متناظر با یک پروتئین و هر یال $e \in E$ بیانگر وجود تعامل میان دو پروتئین است. در حالت کلی، منابع داده PPI را می‌توان به سه دسته شامل داده‌های آزمایشگاهی، پایگاه‌های داده مبتنی بر روش‌های محاسباتی، و پایگاه‌های داده ادغام‌شده تقسیم‌بندی کرد. شبکه‌های PPI برای گونه‌های زیستی مختلفی نظیر انسان، موش و مخمر در دسترس هستند. با این حال، در این پژوهش تمرکز صرفاً بر شبکه‌های مربوط به گونه مخمر نان^۱ قرار دارد؛ چرا که بخش عمده‌ای از داده‌های مرجع، مطالعات پیشین و ماژول‌های عملکردی تأییدشده برای این گونه زیستی فراهم شده‌اند. تمامی مجموعه‌داده‌های مورد استفاده مربوط به این گونه بوده و تفاوت آن‌ها عمدتاً در تعداد گره‌ها و یال‌ها و نیز نوع روش استخراج تعاملات است. از جمله پایگاه‌های داده شناخته‌شده مربوط به گونه مخمر نان می‌توان به BioGRID [۷۴]، DIP [۷۵] و Collins [۷۶] اشاره کرد.

در این پژوهش، شبکه‌های PPI نه تنها به عنوان یک ساختار گرافی، بلکه به عنوان ورودی اصلی مدل‌های یادگیری مبتنی بر شبکه‌های عصبی گرافی در نظر گرفته می‌شوند. از این رو، علاوه بر توپولوژی شبکه، وجود وزن یال‌ها و ویژگی‌های معنایی گره‌ها نقش کلیدی در فرآیند یادگیری و استخراج نمایش‌های نهفته ایفا می‌کند. همچنین، به منظور ارزیابی عملکرد روش پیشنهادی در شناسایی ماژول‌های عملکردی، از مجموعه‌های مرجع شامل ماژول‌های پروتئینی شناخته‌شده نظیر CYC2008 و MIPS [۷۷] به عنوان معیار صحت‌سنجی استفاده می‌شود.

۱.۲.۴ مجموعه‌داده شبکه‌های PPI

شبکه‌های PPI مورد استفاده در این پژوهش به صورت گراف‌های بدون جهت و وزن‌دار مدل‌سازی می‌شوند که به طور رسمی به شکل

$$G = (V, E, W, X)$$

¹ *Saccharomyces cerevisiae*

قابل نمایش هستند. در این نمایش، V مجموعه گره‌ها (پروتئین‌ها)، E مجموعه یال‌ها (تفاعلات پروتئینی)، W ماتریس وزن یال‌ها و X ماتریس ویژگی‌های گره‌ها است. این چارچوب نمایش، مستقیماً با الزامات ورودی شبکه‌های عصبی گرافی مورد استفاده در روش پیشنهادی سازگار است.

مجموعه داده‌های Krogan-Core، Gavin، Collins و Krogan-Extended از صفحه مربوط به الگوریتم IMHRC [۷۸] که توسط پژوهشگاه دانش‌های بنیادی (IPM) منتشر شده است، دریافت شده‌اند. این مجموعه داده‌ها شامل وزن یال‌ها بوده و بنابراین مستقیماً به عنوان گراف‌های وزن‌دار در مدل‌های عصبی گرافی مورد استفاده قرار می‌گیرند.

در مقابل، برای شبکه‌هایی نظیر BioGRID و DIP که فاقد وزن یال هستند، داده‌ها از پژوهش AdaPPI استخراج شده‌اند. از آن‌جا که روش پیشنهادی مبتنی بر یادگیری پیام در شبکه‌های عصبی گرافی نیازمند وزن یال‌ها به منظور تنظیم شدت انتشار اطلاعات میان گره‌ها است، وزن هر یال با استفاده از شباهت کسینوسی بین بردارهای ویژگی مبتنی بر هستی‌شناسی ژن (GO) مربوط به هر جفت پروتئین محاسبه شده است. این بردارها ماتریس ویژگی گره‌ها X را تشکیل داده و نحوه استخراج آن‌ها در بخش بعدی تشریح خواهد شد.

به منظور شفاف‌سازی ساختار داده‌ها، مجموعه داده Collins به عنوان نمونه بررسی می‌شود. هر مجموعه داده به صورت یک فایل جدولی با فرمت CSV ارائه شده و شامل فیلدهای زیر است:

- **protein1**: شناسه یا نام معنایی^۱ پروتئین اول (گره مبدأ).
- **protein2**: شناسه یا نام معنایی پروتئین دوم (گره مقصد).
- **weight**: وزن یال بین دو پروتئین که شدت تعامل یا میزان شباهت زیستی را مدل‌سازی می‌کند.

^۱ Semantic name

۲.۲.۴ استخراج ویژگی‌ها از پایگاه هستی‌شناسی ژن

در این بخش، نحوه استخراج ویژگی‌های زیستی مربوط به هر پروتئین با استفاده از پایگاه داده هستی‌شناسی ژن (Gene Ontology یا GO) تشریح می‌شود. ویژگی‌های استخراج‌شده، ماتریس ویژگی گره‌ها X را در نمایش گرافی شبکه‌های PPI تشکیل داده و به‌عنوان ورودی اصلی شبکه عصبی گرافی در روش پیشنهادی مورد استفاده قرار می‌گیرند.

پایگاه داده GO شامل سه نسخه اصلی go-basic، go و go-plus است که هر یک از نظر سطح جزئیات و نوع روابط موجود میان عبارات تفاوت دارند. در این پژوهش، نسخه go-basic به‌عنوان منبع اصلی استخراج ویژگی‌ها انتخاب شده است؛ چرا که این نسخه شامل روابط سلسله‌مراتبی اصلی بوده و از ایجاد چرخه‌های منطقی جلوگیری می‌کند، که این امر برای استخراج ویژگی‌های پایدار و قابل تفسیر اهمیت دارد. خلاصه‌ای از تفاوت این نسخه‌ها در جدول مربوطه ارائه شده است.

یه جدول طلایی

برای هر گونه‌ی زیستی، یک فایل حاشیه‌نویسی ژن^۱ وجود دارد که نگاشت میان ژن‌ها و عبارات GO را مشخص می‌کند. در این پژوهش، تمرکز بر گونه مخمر نان است و برای هر ژن، شناسه اختصاصی آن موسوم به SGD ID مورد استفاده قرار می‌گیرد. این شناسه امکان استخراج عبارات GO متناظر با هر ژن را در سه زیرهستی‌شناسی شامل فرایند زیستی (Biological Process)، عملکرد مولکولی (Molecular Function) و مؤلفه سلولی (Cellular Component) فراهم می‌کند.

از آن‌جا که در شبکه‌های PPI، پروتئین‌ها اغلب با نام‌های معنایی معرفی می‌شوند، یک گام پیش‌پردازشی برای نگاشت این نام‌ها به SGD ID متناظر آن‌ها ضروری است. بدین منظور، از ابزار ارائه‌شده در وب‌سایت yeastgenome.org استفاده شده است که امکان تبدیل نام‌های معنایی پروتئین‌های مخمر به شناسه‌های SGD را به‌صورت خودکار فراهم می‌کند. پس از انجام این نگاشت، تمامی عبارات GO

¹ Gene Annotation File (GAF)

مرتبط با هر پروتئین از فایل حاشیه نویسی ژن استخراج می‌شوند.

در ساده‌ترین حالت، می‌توان هر پروتئین را به‌صورت یک بردار دودویی با بُعدی برابر تعداد کل عبارات GO نمایش داد، به‌طوری‌که هر مؤلفه نشان‌دهنده وجود یا عدم وجود یک عبارت GO برای آن پروتئین باشد. به‌عنوان نمونه، شبکه PPI مربوط به مجموعه داده Collins شامل ۲۲۴۸ عبارت GO متمایز است و در نتیجه، نمایش دودویی ویژگی‌های پروتئین‌ها دارای همین بُعد خواهد بود. با وجود آن‌که نمایش‌های دودویی مبتنی بر GO به‌دلیل سادگی و تفسیرپذیری، در بسیاری از پژوهش‌های پیشین مورد استفاده قرار گرفته‌اند، این بردارها به‌طور ذاتی تنک و دارای بُعد بالا هستند که می‌تواند منجر به کاهش کارایی یادگیری در مدل‌های مبتنی بر شبکه‌های عصبی گرافی شود.

بر این اساس، در این پژوهش علاوه بر استفاده از بردارهای ویژگی دودویی، از بردارهای تعبیه‌شده حاصل از روش‌های یادگیری بازنمایی گراف، نظیر Node2Vec، نیز بهره گرفته شده است تا اطلاعات معنایی و ساختاری نهفته در روابط میان عبارات GO به‌صورت فشرده‌تری در نمایش ویژگی‌ها منعکس شود.

به‌منظور دستیابی به نمایش‌های غنی‌تر و کم‌بعدتر، از رویکرد تعبیه‌سازی عبارات GO مشابه روش ارائه‌شده در [۷۹] استفاده شده است. بدین منظور، ابتدا الگوریتم Node2Vec بر روی گراف GO اعمال شده و برای هر عبارت GO یک بردار تعبیه‌ای استخراج می‌شود. سپس، برای هر پروتئین و در هر یک از سه زیرهستی‌شناسی فرایند زیستی، عملکرد مولکولی و مؤلفه سلولی، بردارهای تعبیه‌ای عبارات GO متناظر با آن پروتئین میانگین‌گیری می‌شوند. در نتیجه، برای هر زیرهستی‌شناسی یک بردار ویژگی با بُعد ۱۲۸ به‌دست می‌آید.

در نهایت، بردارهای حاصل از سه زیرهستی‌شناسی با یکدیگر الحاق شده و یک بردار ویژگی نهایی با بُعد ۳۸۴ برای هر پروتئین تشکیل می‌شود. این بردار به‌عنوان نمایش ویژگی گره متناظر با هر پروتئین در ماتریس X قرار گرفته و به‌صورت یکپارچه در فرایند یادگیری شبکه عصبی گرافی روش پیشنهادی

مورد استفاده قرار می‌گیرد.

این جا به جدول برای ویژگی‌های دیتاست‌ها

۳.۲.۴ مجموعه‌های پروتئینی مرجع و پروتکل ارزیابی

به منظور ارزیابی و صحت‌سنجی عملکرد الگوریتم پیشنهادی در شناسایی ماژول‌های عملکردی پروتئینی، استفاده از مجموعه‌های پروتئینی مرجع ضروری است تا بتوان نتایج به دست آمده را با دانش زیستی موجود مقایسه کرد. مجموعه‌های پروتئینی مرجع وابسته به گونه زیستی بوده و برای گونه مخمر نان چندین مجموعه مرجع شناخته شده در دسترس است که از جمله مهم‌ترین آن‌ها می‌توان به MIPS و CYC2008 اشاره کرد. در مقایسه روش‌های محاسباتی مختلف، انتخاب مجموعه پروتئینی مرجع مورد استفاده نقش تعیین‌کننده‌ای در تفسیر نتایج ایفا می‌کند. از این رو، با توجه به این که روش پایه مورد استفاده برای مقایسه با روش پیشنهادی، چارچوب AdaPPI است، در این پژوهش نیز از همان مجموعه‌های پروتئینی مرجع به کاررفته در آن مطالعه استفاده شده است. این مجموعه ترکیبی از سایر مجموعه‌های پروتئینی شامل MIPS، CYC2008، Aloy، SGD و TAP06 می‌باشد.

این مجموعه پروتئینی مرجع به صورت یک فایل متنی ذخیره شده است که در آن، هر خط متناظر با یک ماژول پروتئینی بوده و شامل نام‌های معنایی پروتئین‌های عضو آن ماژول است که با خط فاصله از یکدیگر جدا شده‌اند. شکل؟؟ نمونه‌ای از ساختار این مجموعه مرجع را نشان می‌دهد.

تصویر

به منظور ارزیابی مجموعه‌های پروتئینی استخراج شده از هر شبکه PPI، ابتدا تنها پروتئین‌هایی که در شبکه PPI متناظر حضور دارند در هر مجموعه مرجع حفظ می‌شوند. سپس، مجموعه‌هایی که تعداد پروتئین‌های آن‌ها کمتر از سه عضو باشد حذف می‌گردند. این فرآیند پیش‌پردازش، مطابق با پروتکل ارزیابی به کاررفته در روش AdaPPI انجام شده و امکان مقایسه منصفانه و سازگار نتایج روش پیشنهادی

با روش‌های مرجع را فراهم می‌کند.

۳.۴ روش پیشنهادی

در این بخش، چارچوب روش پیشنهادی این پژوهش برای شناسایی ماژول‌های عملکردی پروتئینی در شبکه‌های برهم‌کنش پروتئین-پروتئین (PPI) ارائه می‌شود. همان‌گونه که در فصل بیان مسئله مطرح شد، تمرکز اصلی این پژوهش بر بهره‌گیری مؤثر از شبکه‌های عصبی گرافی در قالب یک چارچوب یادگیری یکپارچه و بدون نظارت به منظور استخراج ماژول‌های عملکردی هم‌پوشان است. در این راستا، تابع هزینه مبتنی بر مدل مولد احتمالی مورد استفاده در پژوهش NOCD به عنوان پایه در نظر گرفته شده و تلاش شده است با بهبود و تطبیق آن با ویژگی‌های زیستی و ساختاری شبکه‌های PPI، عملکرد شناسایی ماژول‌ها ارتقا یابد.

علاوه بر طراحی تابع هزینه مناسب، این پژوهش به بررسی و مقایسه معماری‌های مختلف شبکه‌های عصبی گرافی و نقش آن‌ها در فرآیند یادگیری بازنمایی گره‌ها می‌پردازد. همچنین، تأثیر نوع ویژگی‌های ورودی شامل ویژگی‌های دودویی مبتنی بر عبارات هستی‌شناسی ژن و بردارهای تعبیه‌شده GO، به صورت نظام‌مند مورد مطالعه قرار گرفته است. در نهایت، یک روش ترکیبی پیشنهاد می‌شود که به طور هم‌زمان از هر دو نوع ویژگی بهره می‌گیرد و هدف آن ارائه یک چارچوب جامع برای شناسایی دقیق‌تر ماژول‌های عملکردی پروتئینی است. در ادامه، ابتدا چارچوب کلی روش پیشنهادی و نحوه توسعه آن بر پایه مدل NOCD تشریح می‌شود. سپس، معماری‌های مختلف شبکه‌های عصبی گرافی مورد استفاده معرفی و تحلیل می‌شوند. در گام بعد، تأثیر نوع ویژگی‌های ورودی بر عملکرد مدل بررسی شده و در پایان، روش ترکیبی پیشنهادی ارائه می‌گردد.

کتاب نامه

- [1] Ji, Junzhong, Zhang, Aidong, Liu, Chunnian, Quan, Xiaomei, and Liu, Zhijun. Survey: Functional module detection from protein-protein interaction networks. *IEEE Transactions on Knowledge and Data Engineering*, 26(2):261–277, 2012.
- [2] V, Manila M. A literature survey on bioinformatics. *IJIREEICE International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, February 2023.
- [3] Wang, Yijie and Qian, Xiaoning. Functional module identification in protein interaction networks by interaction patterns. *Bioinformatics*, 30(1):81–93, 2014.
- [4] Berahmand, Kamal, Nasiri, Elahe, Li, Yuefeng, et al. Spectral clustering on protein-protein interaction networks via constructing affinity matrix using attributed graph embedding. *Computers in Biology and Medicine*, 138:104933, 2021.
- [5] Li, Xiaoli, Wu, Min, Kwoh, Chee-Keong, and Ng, See-Kiong. Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC genomics*, 11:1–19, 2010.

- [6] Bader, Gary D and Hogue, Christopher WV. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, 4:1–27, 2003.
- [7] Nepusz, Tamás, Yu, Haiyuan, and Paccanaro, Alberto. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*, 9(5):471–472, 2012.
- [8] Consortium, Gene Ontology. The gene ontology (go) project in 2006. *Nucleic acids research*, 34(suppl_1):D322–D326, 2006.
- [9] Su, Lili, Liu, Guang, Guo, Ying, Zhang, Xuanping, Zhu, Xiaoyan, and Wang, Jiayin. Integration of protein-protein interaction networks and gene expression profiles helps detect pancreatic adenocarcinoma candidate genes. *Frontiers in Genetics*, 13:854661, 2022.
- [10] Bothorel, Cécile, Cruz, Juan David, Magnani, Matteo, and Micenkova, Barbora. Clustering attributed graphs: models, measures and methods. *Network Science*, 3(3):408–444, 2015.
- [11] Farutin, Victor, Robison, Keith, Lightcap, Eric, Dancik, Vlado, Ruttenberg, Alan, Letovsky, Stanley, and Pradines, Joel. Edge-count probabilities for the identification of local protein communities and their organization. *Proteins: Structure, Function, and Bioinformatics*, 62(3):800–818, 2006.
- [12] Altaf-Ul-Amin, Md, Shinbo, Yoko, Mihara, Kenji, Kurokawa, Ken, and Kanaya, Shigehiko. Development and implementation of an algorithm for de-

tection of protein complexes in large interaction networks. *BMC bioinformatics*, 7(1):207, 2006.

- [13] Zaki, Nazar and Alashwal, Hany. Improving the detection of protein complexes by predicting novel missing interactome links in the protein-protein interaction network. in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5041–5044. IEEE, 2018.
- [14] Macropol, Kathy, Can, Tolga, and Singh, Ambuj K. Rrw: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC bioinformatics*, 10(1):283, 2009.
- [15] Chen, Hongwei, Cai, Yunpeng, Ji, Chaojie, Selvaraj, Gurudeeban, Wei, Dongqing, and Wu, Hongyan. Adappi: identification of novel protein functional modules via adaptive graph convolution networks in a protein–protein interaction network. *Briefings in Bioinformatics*, 24(1):bbac523, 2023.
- [16] Alberts, Bruce, Heald, Rebecca, Johnson, Alexander, Morgan, David, Raff, Martin, Roberts, Keith, and Walter, Peter. *Molecular Biology of the Cell*. W. W. Norton & Company, New York, 7th ed. , 2022. International Student Edition.
- [17] Dilmaghani, Saharnaz, Brust, Matthias R, Ribeiro, Carlos HC, Kieffer, Emmanuel, Danoy, Grégoire, and Bouvry, Pascal. From communities to protein complexes: a local community detection algorithm on ppi networks. *Plos one*, 17(1):e0260484, 2022.
- [18] Hartwell, Leland H, Hopfield, John J, Leibler, Stanislas, and Murray, An-

drew W. From molecular to modular cell biology. *Nature*, 402(Suppl 6761):C47–C52, 1999.

- [19] Li, Min, Wu, Xuehong, Wang, Jianxin, and Pan, Yi. Towards the identification of protein complexes and functional modules by integrating ppi network and gene expression data. *BMC bioinformatics*, 13(1):109, 2012.
- [20] Safari-Alighiarloo, Nahid, Taghizadeh, Mohammad, Rezaei-Tavirani, Mostafa, Goliaei, Bahram, and Peyvandi, Ali Asghar. Protein-protein interaction networks (ppi) and complex diseases. *Gastroenterology and Hepatology from bed to bench*, 7(1):17, 2014.
- [21] Mujawar, Shama, Mishra, Rohit, Pawar, Shrikant, Gatherer, Derek, and Lahiri, Chandrajit. Delineating the plausible molecular vaccine candidates and drug targets of multidrug-resistant acinetobacter baumannii. *Frontiers in cellular and infection microbiology*, 9:203, 2019.
- [22] Gélard, Maxence, Richard, Guillaume, Pierrot, Thomas, and Cournède, Paul-Henry. Bulkcrabert: Cancer prognosis from bulk rna-seq based language models. *bioRxiv*, pp. 2024–06, 2024.
- [23] Consortium, Gene Ontology. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338, 2019.
- [24] Wikipedia contributors. ژن مس‌تیشناسی — Wikipedia, the free encyclopedia, 2024. [Online; accessed 20-December-2024].
- [25] Gene Ontology overview, December 2025. [Online; accessed 17. Dec. 2025].

- [26] Ashburner, Michael, Ball, Catherine A, Blake, Judith A, Botstein, David, Butler, Heather, Cherry, J Michael, Davis, Allan P, Dolinski, Kara, Dwight, Selina S, Eppig, Janan T, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.
- [27] Stumpf, Michael PH, Kelly, William P, Thorne, Thomas, and Wiuf, Carsten. Evolution at the system level: the natural history of protein interaction networks. *Trends in Ecology & Evolution*, 22(7):366–373, 2007.
- [28] Li, Dong, Li, Jianqi, Ouyang, Shuguang, Wang, Jian, Wu, Songfeng, Wan, Ping, Zhu, Yunping, Xu, Xiaojie, and He, Fuchu. Protein interaction networks of *saccharomyces cerevisiae*, *caenorhabditis elegans* and *drosophila melanogaster*: Large-scale organization and robustness. *Proteomics*, 6(2):456–461, 2006.
- [29] Hartwell, Leland H, Hopfield, John J, Leibler, Stanislas, and Murray, Andrew W. From molecular to modular cell biology. *Nature*, 402(Suppl 6761):C47–C52, 1999.
- [30] Wagner, Günter P, Pavlicev, Mihaela, and Cheverud, James M. The road to modularity. *Nature Reviews Genetics*, 8(12):921–931, 2007.
- [31] Islam, Rakibul, Sultana, Azrin, and Islam, Mohammad Rashedul. A comprehensive review for chronic disease prediction using machine learning algorithms. *Journal of Electrical Systems and Information Technology*, 11(1):27, 2024.
- [32] Bishop, Christopher M and Nasrabadi, Nasser M. *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

- [33] Janiesch, Christian, Zschech, Patrick, and Heinrich, Kai. Machine learning and deep learning. *Electronic markets*, 31(3):685–695, 2021.
- [34] Dridi, Salim. Supervised learning-a systematic literature review. *preprint, Dec*, 2021.
- [35] Wu, Xiangdong, Liu, Xiaoyan, and Zhou, Yimin. Review of unsupervised learning techniques. in *Proceedings of 2021 Chinese Intelligent Systems Conference: Volume II*, pp. 576–590. Springer, 2021.
- Tehran, Mobtakeran. *Mathematics Discrete in Topics* Esmail. Babalian. [۳۶]
۲۰۰۷
- [37] Scarselli, Franco, Gori, Marco, Tsoi, Ah Chung, Hagenbuchner, Markus, and Monfardini, Gabriele. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.
- [38] Grattarola, Daniele. A practical introduction to GNNs - Part 2: Message passing and gather-scatter. Daniele Grattarola’s Blog, March 2021. Accessed: 2024-12-20.
- [39] Rong, Yu, Huang, Wenbing, Xu, Tingyang, and Huang, Junzhou. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*, 2019.
- [40] Khemani, Bharti, Patil, Shruti, Kotecha, Ketan, and Tanwar, Sudeep. A review of graph neural networks: concepts, architectures, techniques, challenges,

datasets, applications, and future directions. *Journal of Big Data*, 11(1):18, 2024.

- [41] Kipf, TN. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- [42] Veličković, Petar, Cucurull, Guillem, Casanova, Arantxa, Romero, Adriana, Lio, Pietro, and Bengio, Yoshua. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- [43] Xu, Keyulu, Li, Chengtao, Tian, Yonglong, Sonobe, Tomohiro, Kawarabayashi, Ken-ichi, and Jegelka, Stefanie. Representation learning on graphs with jumping knowledge networks. in *International conference on machine learning*, pp. 5453–5462. pmlr, 2018.
- [44] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [45] Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [46] Grover, Aditya and Leskovec, Jure. node2vec: Scalable feature learning for networks. in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864, 2016.

- [47] Jannesari, Vahid, Keshvari, Maryam, and Berahmand, Kamal. A novel non-negative matrix factorization-based model for attributed graph clustering by incorporating complementary information. *Expert Systems with Applications*, 242:122799, 2024.
- [48] Kang, Zhao, Liu, Zhanyu, Pan, Shirui, and Tian, Ling. Fine-grained attributed graph clustering. in *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pp. 370–378. SIAM, 2022.
- [49] Zhang, Xiaotong, Liu, Han, Li, Qimai, Wu, Xiao-Ming, and Zhang, Xianchao. Adaptive graph convolution methods for attributed graph clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12384–12399, 2023.
- [50] Bhowmick, Aritra, Kosan, Mert, Huang, Zexi, Singh, Ambuj, and Medya, Sourav. Dgcluster: A neural framework for attributed graph clustering via modularity maximization. in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 11069–11077, 2024.
- [51] Zhang, Tian, Ramakrishnan, Raghu, and Livny, Miron. Birch: an efficient data clustering method for very large databases. *ACM sigmod record*, 25(2):103–114, 1996.
- [52] He, Chaobo, Cheng, Junwei, Chen, Guohua, Guan, Quanlong, Fei, Xiang, and Tang, Yong. Detecting communities with multiple topics in attributed networks via self-supervised adaptive graph convolutional network. *Information Fusion*, 105:102254, 2024.

- [53] Wang, Chun, Pan, Shirui, Hu, Ruiqi, Long, Guodong, Jiang, Jing, and Zhang, Chengqi. Attributed graph clustering: A deep attentional embedding approach. *arXiv preprint arXiv:1906.06532*, 2019.
- [54] Shchur, Oleksandr and Günnemann, Stephan. Overlapping community detection with graph neural networks. *arXiv preprint arXiv:1909.12201*, 2019.
- [55] Srihari, Sriganesh and Leong, Hon Wai. Employing functional interactions for characterisation and detection of sparse complexes from yeast ppi networks. *International journal of bioinformatics research and applications*, 8(3-4):286–304, 2012.
- [56] Patra, Sabyasachi and Mohapatra, Anjali. Protein complex prediction in interaction network based on network motif. *Computational Biology and Chemistry*, 89:107399, 2020.
- [57] Nepusz, Tamás, Yu, Haiyuan, and Paccanaro, Alberto. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*, 9(5):471–472, 2012.
- [58] Bader, Gary D and Hogue, Christopher WV. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, 4:1–27, 2003.
- [59] Hu, Allen L and Chan, Keith CC. Utilizing both topological and attribute information for protein complex identification in ppi networks. *IEEE/ACM transactions on computational biology and bioinformatics*, 10(3):780–792, 2013.

- [60] Zhang, Wei, Xu, Jia, Li, Yuanyuan, and Zou, Xiufen. Integrating network topology, gene expression data and go annotation information for protein complex prediction. *Journal of bioinformatics and computational biology*, 17(01):1950001, 2019.
- [61] Wu, Min, Li, Xiaoli, Kwoh, Chee-Keong, and Ng, See-Kiong. A core-attachment based method to detect protein complexes in ppi networks. *BMC bioinformatics*, 10:1–16, 2009.
- [62] Leung, Henry CM, Xiang, Qian, Yiu, Siu-Ming, and Chin, Francis YL. Predicting protein complexes from ppi data: a core-attachment approach. *Journal of Computational Biology*, 16(2):133–144, 2009.
- [63] Chua, Hon Nian, Ning, Kang, Sung, Wing-Kin, Leong, Hon Wai, and Wong, Limsoon. Using indirect protein–protein interactions for protein complex prediction. *Journal of bioinformatics and computational biology*, 6(03):435–466, 2008.
- [64] Berahmand, Kamal, Nasiri, Elahe, Li, Yuefeng, et al. Spectral clustering on protein-protein interaction networks via constructing affinity matrix using attributed graph embedding. *Computers in Biology and Medicine*, 138:104933, 2021.
- [65] Yang, Cheng, Liu, Zhiyuan, Zhao, Deli, Sun, Maosong, and Chang, Edward Y. Network representation learning with rich text information. in *IJCAI*, vol. 2015, pp. 2111–2117, 2015.

- [66] Meng, Xiangmao, Peng, Xiaoqing, Wu, Fang-Xiang, and Li, Min. Detecting protein complex based on hierarchical compressing network embedding. in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 215–218. IEEE, 2019.
- [67] Xu, Bo, Li, Kun, Zheng, Wei, Liu, Xiaoxia, Zhang, Yijia, Zhao, Zhehuan, and He, Zengyou. Protein complexes identification based on go attributed network embedding. *BMC bioinformatics*, 19:1–10, 2018.
- [68] Ma, Xiaoke, Sun, Penggang, and Gong, Maoguo. An integrative framework of heterogeneous genomic data for cancer dynamic modules based on matrix decomposition. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(1):305–316, 2020.
- [69] Chen, Hongwei, Cai, Yunpeng, Ji, Chaojie, Selvaraj, Gurudeeban, Wei, Dongqing, and Wu, Hongyan. Adappi: identification of novel protein functional modules via adaptive graph convolution networks in a protein–protein interaction network. *Briefings in Bioinformatics*, 24(1):bbac523, 2023.
- [70] Bhowmick, Sourav S and Seah, Boon Siew. Clustering and summarizing protein-protein interaction networks: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):638–658, 2015.
- [71] Berahmand, Kamal, Bouyer, Asgarali, and Vasighi, Mahdi. Community detection in complex networks by detecting and expanding core nodes through extended local similarity of nodes. *IEEE Transactions on Computational Social*

Systems, 5(4):1021–1033, 2018.

- [72] Zhou, Zhixin and Amini, Arash A. Analysis of spectral clustering algorithms for community detection: the general bipartite setting. *Journal of Machine Learning Research*, 20(47):1–47, 2019.
- [73] Gulikers, Lennart, Lelarge, Marc, and Massoulié, Laurent. A spectral method for community detection in moderately sparse degree-corrected stochastic block models. *Advances in Applied Probability*, 49(3):686–721, 2017.
- [74] Stark, Chris, Breitkreutz, Bobby-Joe, Reguly, Teresa, Boucher, Lorrie, Breitkreutz, Ashton, and Tyers, Mike. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl_1):D535–D539, 2006.
- [75] Xenarios, Ioannis, Salwinski, Lukasz, Duan, Xiaoqun Joyce, Higney, Patrick, Kim, Sul-Min, and Eisenberg, David. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic acids research*, 30(1):303–305, 2002.
- [76] Collins, Sean R, Kemmeren, Patrick, Zhao, Xue-Chu, Greenblatt, Jack F, Spencer, Forrest, Holstege, Frank CP, Weissman, Jonathan S, and Krogan, Nevan J. Toward a comprehensive atlas of the physical interactome of *saccharomyces cerevisiae*. *Molecular & Cellular Proteomics*, 6(3):439–450, 2007.
- [77] Pagel, Philipp, Kovac, Stefan, Oesterheld, Matthias, Brauner, Barbara, Dunger-Kaltenbach, Irmtraud, Frishman, Goar, Montrone, Corinna, Mark, Pekka,

Stümpflen, Volker, Mewes, Hans-Werner, et al. The mips mammalian protein–protein interaction database. *Bioinformatics*, 21(6):832–834, 2005.

[78] Maddi, AMA and Eslahchi, Ch. Discovering overlapped protein complexes from weighted ppi networks by removing inter-module hubs. *Scientific reports*, 7(1):3247, 2017.

[79] Ieremie, Ioan, Ewing, Rob M, and Niranjan, Mahesan. TransformerGO: predicting protein–protein interactions by modelling the attention between sets of gene ontology terms. *Bioinformatics*, 38(8):2269–2277, 2022.

واژه‌نامه انگلیسی به فارسی

Example مثال

module مدول

واژه‌نامه فارسی به انگلیسی

Example مثال

module مدول

Abstract

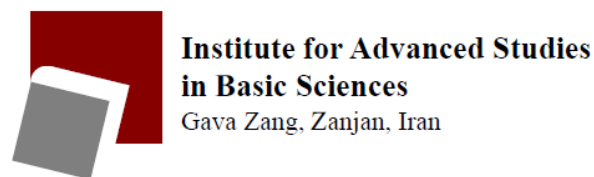
Bioinformatics is an interdisciplinary field that utilizes biology, computer science, mathematics, and statistics to store and analyze biological data. With the completion of the Human Genome Project and the advent of the post-genomic era, proteomics research has become one of the most important areas of life sciences. Proteomics involves studying the characteristics of proteins to describe their structure, function, and role in regulating biological systems. Proteins often do not act alone but interact with each other, forming larger molecular complexes to perform biological functions. These interactions are represented using a network structure called the protein-protein interaction (PPI) network. A protein complex in PPI networks is a molecular structure composed of proteins that are functionally and structurally compatible. By analyzing PPI networks, we can identify these groups of proteins.

One of the key challenges in bioinformatics is the discovery of protein modules in protein-protein interaction networks. Identifying these modules is equivalent to the problem of community detection in graphs. In many bioinformatics applications, protein module discovery is performed using community detection algorithms in graphs. In this study, we aim to design a specialized method for community detection in protein interaction networks that, in addition to considering the graph structure for module identification, also takes into account the biological characteristics of proteins.

For example, integrating biological information about proteins stored in databases such as GO and KEGG with gene expression data and combining this information with

the PPI network can enhance the accuracy and efficiency of protein module identification. Therefore, in this research, we aim to introduce a clustering algorithm for PPI networks based on graph neural networks while incorporating node-specific features.

Keywords: *Graph Neural Networks, Protein-Protein Interactions, Functional Module Identification, Clustering Attributed Graphs*



Computer Science and Information Technology
Artificial Intelligence

Discovery of Modules in Protein-Protein Interaction Networks using Graph Neural Network Approaches

Master's Thesis

Samaneh Tejerloo

Supervisor: Dr. Zahra Narimani

January 7, 2026