



دانشکده علوم رایانه و فناوری اطلاعات

مهندسی کامپیوتر - هوش مصنوعی

کشف ماذولها در شبکه‌های پروتئین – پروتئین با رویکردهای شبکه‌های عصبی گرافی

پایاننامه‌ی کارشناسی ارشد

سمانه طجرلو

استاد راهنما: زهرا نریمانی

بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِيْمِ

تقدیم به آنها یک که می خواهند بیشتر بدانند.

مشکر 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 پایان یافتن پژوهه توالی‌یابی ژنوم انسان و ورود به دوره‌ی پسازنی، 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 ماذول‌های پروتئین در شبکه‌های تعامل پروتئین-پروتئین است. کشف این ماذول‌ها معادل مسئله‌ی کشف انجمن در گراف است. در بسیاری از کاربردهای بیوانفورماتیکی کشف ماذول‌های 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 ماذول‌ها 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 ماذول‌های 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 مأذول‌های عملکردی، خوشه‌بندی گراف‌های دارای ویژگی

فهرست مطالب

پنج	چکیده
۱	پیشگفتار
۲	۱ معرفی پژوهش
۲	۱.۱ مقدمه
۳	۲.۱ بیان مسئله
۶	۳.۱ اهمیت و ضرورت انجام پژوهش
۶	۴.۱ پرسش‌های پژوهش
۷	۵.۱ روش پژوهش
۷	۶.۱ جمع‌بندی
۹	۲ مفاهیم بنیادی
۹	۱.۲ مقدمه
۱۰	۲.۲ مفاهیم زیستی
۱۰	۱.۲.۲ پروتئین
۱۱	۲.۲.۲ مازول‌های عملکردی
۱۲	۳.۲.۲ بیان زن

۱۲	پایگاه داده هستی شناسی ژن	۴.۲.۲
۱۴	شبکه‌های PPI و ویژگی‌های آنها	۵.۲.۲
۱۵	مفاهیم محاسباتی	۳.۲
۱۵	یادگیری ماشین	۱.۲.۲
۱۶	یادگیری عمیق	۲.۲.۲
۱۶	یادگیری ناظارت شده	۳.۲.۲
۱۷	یادگیری بدون ناظارت	۴.۲.۲
۱۷	گراف	۵.۲.۲
۱۸	شبکه‌های عصبی گرافی	۶.۲.۲
۱۹	شبکه‌های عصبی گرافی پیچشی	۷.۲.۲
۲۰	شبکه‌های عصبی گرافی توجه محور	۸.۲.۲
۲۲	شبکه‌های عصبی گرافی با انتقال دانش بین لایه‌ها	۹.۲.۲
۲۴	تعبیه گره‌ها به روش Node2Vec	۱۰.۲.۲
۲۵	خوشه‌بندی در گراف‌های با گره‌های دارای ویژگی	۱۱.۲.۲
۲۵	دسته‌بندی و روش‌های کلی خوشه‌بندی گراف	۱۲.۲.۲
۲۶	معیارهای ارزیابی	۴.۲
۲۷	شباهت همسایگی	۱.۴.۲
۲۸	دقت	۲.۴.۲
۲۸	بازیابی	۳.۴.۲
۲۸	امتیاز F	۴.۴.۲
۲۹	صحت	۵.۴.۲

۳۰	خوشبندی گرافهای دارای گره ویژگی	۱.۳
۳۳	پیش‌بینی مجموعه‌های پروتئینی	۲.۳
۳۷		روش ۴
۳۷	مجموعه داده	۱.۴
۳۸	روش پیشنهادی	۲.۴
۳۸	مرحله اول: استفاده از شبکه عصبی گرافی به منظور ایجاد ماتریس وابستگی	۱.۲.۴
۳۹	مرحله دوم: بهینه‌سازی وزن‌های شبکه عصبی گرافی	۲.۲.۴
۴۰	مرحله سوم: تخصیص نودها به خوشه‌ها	۳.۲.۴
۵۲	واژه‌نامه انگلیسی به فارسی	
۵۳	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 کردایم. 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

بیانفورماتیک سعی بر پاسخ به مسائل و پرسش‌های زیست شناسی به کمک ابزارهای محاسباتی و مدل‌سازی‌های آماری دارد. یافتن پاسخ مناسب برای هر یک از این مسائل می‌تواند تاثیر به سزاگی 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 تعاملات پروتئین-پروتئین می‌باشد.

پروتئین‌ها مولکول‌های بزرگ و پیچیده‌ای هستند که وظایف زیستی حیاتی‌ای مانند نقش ساختاری و حمایتی از سلول‌ها، عملکرد پادتنی، نقش‌های پیام رسانی و یا نقش آنزیمی را عهده دار هستند. بسیاری از فرآیندهای زیستی به وسیله مجموعه‌های از پروتئین‌ها انجام می‌گیرد که مأذول‌های عملکردی نامیده می‌شوند. شناخت هر چه بهتر این مجموعه‌های پروتئینی به درک بهتر ما از فرآیندهای زیستی و همچین درک نقش پروتئین‌های کمتر شناخته شده کمک شایانی می‌کند. تشخیص این مأذول‌ها 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.

۲.۱ بیان مسئله^۱

بیوانفورماتیک^۲ یک حوزه میانرشته‌ای است که با استفاده از علوم زیست‌شناسی، کامپیوتر، ریاضیات و آمار به ذخیره‌سازی و تحلیل داده‌های زیستی می‌پردازد. این علم با بهره‌گیری از فناوری‌های کامپیوتری، داده‌های مربوط به توالی‌های دی‌ان‌ای^۳، آرزنای^۴ و پروتئین‌ها را مدیریت و تفسیر می‌کند و به دلیل حجم بالای داده‌ها و اهمیت استخراج اطلاعات کاربردی، جایگاه ویژه‌ای دارد [۱].

با پایان یافتن پژوهه توالی‌یابی ژنوم انسان و ورود به دوره‌ی پسازنی^۵، تحقیقات پروتئومیک^۶ 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 پروتئین-پروتئین^۷ در فرآیندهای مهمی مانند تکثیر ژنتیکی^۸، کنترل بیان ژن^۹، انتقال سیگنال‌های سلولی^{۱۰} و مرگ سلولی^{۱۱} نقش کلیدی دارند. تحلیل شبکه‌های PPI برای درک بهتر سازماندهی و عملکرد سلولی ضروری است [۲].

Problem statement	^۱
Bioinformatic	^۲
DNA	^۳
RNA	^۴
Postgenomic era	^۵
Proteomics	^۶
Protein-protein interactions (PPI)	^۷
Gene substance copy	^۸
Gene expression control	^۹
Cellular signal transduction	^{۱۰}
Cell apoptosis	^{۱۱}

تحقیقات زیستی نشان می‌دهد که یک مجموعه پروتئینی^۱ 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 [۲]. به صورت شهودی نیز، در شبکه‌های PPI اگر دو پروتئین با هم تعامل داشته باشند، به احتمال بیشتری از نظر کارایی سلولی نیز شبیه به یکدیگر هستند. از این رو پیدا کردن زیرشبکه‌هایی به هم متصل با تراکم بالا از پروتئین‌ها می‌تواند به عنوان ماذول‌های عملکردی^۲ و یا یک مجموعه پروتئینی در نظر گرفته شوند که در فرآیندهای ویژه‌ای در سلول نقش دارند [۳]. همینطور بر این اساس می‌توان تعامل و ارتباط بین مجموعه‌های پروتئینی مختلف را بررسی کرد و یا حتی مجموعه‌های پروتئینی ناشناخته‌ای را کشف کرد [۴].

یکی از بهترین روش‌ها برای مطالعه شبکه‌های PPI نگاه به آن‌ها از دید تحلیل شبکه‌های پیچیده و دید گرافی است. با این دید و به دلیل حجم عظیم داده‌های این شبکه‌ها، یکی از چالش‌های اساسی در دوره پسازنی، ارائه الگوریتم‌های بهینه به منظور شناسایی مؤثر ماذول‌های عملکردی زیستی و مجموعه‌های پروتئینی است [۴]. از آنجایی که پروتئین‌های یک مجموعه پروتئینی در شبکه PPI دارای تعاملات زیادی بین خودشان هستند و این موضوع باعث کارکرد مشابه آن‌ها می‌شود، در نتیجه نواحی پر تراکم در شبکه PPI را می‌توان به عنوان یک مجموعه پروتئینی احتمالی در نظر گرفت و شناسایی مجموعه‌های پروتئینی بسیار شبیه به پیدا کردن خوشه‌ها در یک شبکه پیچیده است [۵]، از این رو می‌توان این مسئله را معادل خوشه‌بندی در گراف‌ها در نظر گرفت.

بیشتر پژوهش‌های پیشین در این حوزه تنها از اطلاعات ساختاری شبکه‌های PPI [۶]، [۷] (یه دو تا مقاله دیگه هم اضافه بشه) استفاده کرده‌اند در حالی که امروزه ما به داده‌های غنی و مناسبی برای توصیف پروتئین‌ها دسترسی داریم. به عنوان مثال می‌توان به بانک اطلاعاتی GO^۳ اشاره کرد که اطلاعاتی درباره ژن‌ها و پروتئین‌ها از دیدهای مختلف شامل فرآیندهای زیستی، عملکرد مولکولی و مولفه‌های

^۱ Protein complex

^۲ Functional module

^۳ Gene Ontology

سلولی فراهم کرده و مورد بررسی و توصیف قرار می‌دهد [۸]. همچنین از آنجایی که پروتئین‌ها محصولات ثانی هستند برخی از پژوهش‌ها از بیان ثانی مربوط به هر پروتئین نیز به منظور توصیف آن‌ها در شبکه PPI بهره برده‌اند [۹]. استفاده از این اطلاعات در کنار ساختار شبکه PPI می‌تواند منجر به تشخیص بهینه و موثرتر مجموعه‌های پروتئینی شود. برای استفاده مناسب از اطلاعات تکمیلی نیاز به الگوریتم‌های خوشبندی گرافی داریم که به طور مناسب ویژگی‌های پروتئین‌ها را در کنار ساختار شبکه برای خوشبندی مجموعه‌های پروتئینی در نظر گیرد، به این دسته از الگوریتم‌ها، خوشبندی گراف‌های دارای ویژگی^۱ می‌گویند [۱۰]. آز آنجایی که یک پروتئین می‌تواند در چندین فرآیند زیستی نقش داشته باشد، در نتیجه بسیاری از مجموعه‌های پروتئین‌ها مشترک هستند. بنابراین الگوریتم پیشنهادی باید توانایی شناسایی مجموعه‌های همپوشان را نیز داشته باشد.

بسیاری از پژوهش‌های پیشین تنها با توجه به ساختار شبکه‌های PPI و بدون در نظر گرفتن ویژگی‌های شناخته شده پروتئین‌ها اقدام به ارائه الگوریتم‌های کلاسیک به منظور شناسایی مأذول‌های عملکردی کرده‌اند. با توجه به پیشرفت‌های اخیر در زمینه هوش مصنوعی و الگوریتم‌های یادگیری ماشین به ویژه الگوریتم‌های یادگیری عمیق نیاز به مطالعه عملکرد شبکه‌های عصبی به ویژگی شبکه‌های عصبی گرافی در این زمینه احساس می‌شود. همینطور برخی از پژوهش‌های پیشین نیز از ترکیب روش‌های کلاسیک با شبکه‌های عصبی بهره برده‌اند ولی روشی صرفاً مبتنی بر یادگیری عمیق و به صورت یکپارچه^۲ ارائه نشده است.

به طور خلاصه، ما سعی در ارائه یک الگوریتم جهت کشف مجموعه‌های پروتئینی و مأذول‌های عملکردی هم پوشان با دید گرافی به شبکه‌های PPI در کنار استفاده از داده‌های تکمیلی بانک‌های داده‌های زیستی، با روش خوشبندی گراف با گره‌های ویژگی دار را داریم. در ادامه اهمیت موضوع، پرسش‌هایی که در این پژوهش به دنبال جواب آن‌ها هستیم و همچنین روش پیشنهادی را به صورت

Attributed graph clustering ^۱
End-to-end ^۲

مختصر شرح خواهیم داد.

۳.۱ اهمیت و ضرورت انجام پژوهش

در حال حاضر زیست شناسان 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 دی‌إن‌آی 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 فرایند 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 زمان بر و هزینه‌بر بودن فرآیند کشف آن‌ها در آزمایشگاه است که سبب توجه بیشتر محققان به روش‌های محاسباتی جهت کشف ماذول‌های عملکردی شده است [۱۵].

۴.۱ پرسش‌های پژوهش

بعد از تکمیل قسمت روش

- الگوریتم‌های موجود به ویژه الگوریتم‌های مبتنی بر GNN برای یافتن انجمن در گراف تا چه حد موفق به کشف شبکه‌های تعامل پروتئین هستند؟
- چگونه می‌توان اطلاعات زیستی را به یک الگوریتم موجود برای بهبود کشف مازول‌های تعاملات پروتئینی اضافه کرد؟
- ترکیب اطلاعات توپولوژی و اطلاعات زیستی تا چه حد کشف مازول‌های پروتئینی را بهبود می‌دهد؟
- وجود همپوشانی بین مازول‌ها در افزایش دقت در کشف مازول‌های پروتئینی تا چه حد اثرگذار است؟

۵.۱ روشن پژوهش

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 مازول‌های 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 دنبال‌کردن 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 ماذول‌های عملکردی معرفی شده و به طور 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 به شمار می‌آیند که از تکرار واحد‌های اسید آمینه ساخته شده‌اند و نقش‌های حیاتی و متنوعی را در ساختار و عملکرد سلول ایفا می‌کنند. این مولکول‌ها به طور ویژه به عنوان کارگزاران مولکولی سلول شناخته می‌شوند، زیرا در فرایندهایی مانند کاتالیز واکنش‌های شیمیایی (از طریق آنزیم‌ها)، انتقال مولکول‌ها، پیام‌رسانی درون‌سلولی، ایجاد حرکت و حفظ یکپارچگی ساختار سلولی نقش اساسی دارند. توالي اسیدهای آمینه هر پروتئین توسط ژن مربوطه تعیین شده و طی فرایند ترجمه از روی آران‌ای (ریبونوکلئیک اسید)^۱ پیام‌رسان سنتز می‌شود. پس از سنتز، پروتئین‌ها از طریق فرایند 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 فرایندهای زیستی سلول، به صورت مستقیم یا غیرمستقیم، تحت تأثیر یا کنترل آن‌ها قرار دارند [۱۶].

mRNA: messenger ribonucleic acid ^۱
Folding ^۲

۲.۲.۲ ماذول‌های 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 به صورت 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].

به چنین مجموعه‌ای از پروتئین‌ها که به صورت هماهنگ برای انجام یک عملکرد زیستی مشترک عمل می‌کنند، مجموعه‌ی پروتئینی یا ماذول عملکردی^۱ گفته می‌شود. هر ماذول عملکردی معمولاً بیانگر یک فرآیند زیستی، مسیر مولکولی یا سازوکار تنظیمی خاص در سلول است و اجزای آن، از نظر عملکردی به یکدیگر وابسته‌اند [۱۶].

تعامل فیزیکی میان پروتئین‌ها که تحت عنوان تعامل پروتئین–پروتئین شناخته می‌شود، نقش محوری در شکل‌گیری و پایداری این ماذول‌های 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 [۱۷].

از جمله فرآیندهای زیستی مهمی که مبتنی بر ماذول‌های عملکردی هستند می‌توان به رونوشت دی‌إن‌ای، رونوشت آر ان‌ای 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 ماذول‌های عملکردی، به یکی از موضوعات مهم در زیست‌شناسی سامانه‌ای و بیوانفورماتیک تبدیل شده است. شناسایی دقیق این ماذول‌ها کاربردهای گسترده‌ای از جمله پیش‌بینی عملکرد پروتئین‌های ناشناخته [۱۹]، درک مکانیسم‌های مولکولی بیماری‌ها

[۲۰] و کشف اهداف دارویی جدید [۲۱] دارد. از این رو، مطالعه و مدل‌سازی مأذول‌های عملکردی نقش کلیدی در توسعه روش‌های نوین تشخیصی و درمانی ایفا می‌کند.

۳.۲.۲ بیان ژن^۱

بیان ژن فرآیندی است که طی آن اطلاعات نهفته در توالی دی‌إن‌ای به محصولات عملکردی، عمدتاً آران‌ای و پروتئین، تبدیل می‌شود. این فرآیند شامل مراحل متعددی از جمله رونوشت دی‌إن‌ای 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 نشان‌دهنده 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 این رو، تحلیل داده‌های بیان ژن ابزار مهمی برای درک فرآیندهای زیستی، شناسایی مسیرهای مولکولی مختلف شده در بیماری‌ها و استخراج نشانگرهای زیستی به‌شمار می‌رود [۲۲].

۴.۲.۲ پایگاه داده هستی‌شناسی ژن^۲

GO 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 پرداش از فرآیندهای زیستی، عملکرد مولکولی و مولفه‌ی 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].

Gene expression ^۱
Gene Ontology ^۲

هر عبارت GO شامل موارد زیر می‌شود [۲۴]:

- 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. گراف GO شامل چهار نوع یال is_a، part_of، regulates و has_part است که هر یک به ترتیب بیانگر رابطه «نوعی از»، «جزئی از»، «نقش تنظیم‌کنندگی» و «دارا بودن جزء» میان مفاهیم مختلف در این هستی‌شناسی می‌باشند [۲۵]. 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 نقش دارد.

عملکرد مولکولی^۳: این بخش عملکرد دقیق مولکولی ژن یا پروتئین را توصیف می‌کند.

مولفه‌ی سلولی^۴: 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 مرتبی است که شامل شبکه‌های بدون دور می‌شود و ویژگی‌ها به این صورت مرتب شده‌اند [۸].

graph acyclic Directed^۱
Biological process^۲
Molecular function^۳
Cellular component^۴

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^۱: $P(k)$ 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 $\sim k^{-\lambda}$ می‌باشد [۲۶]. 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 گره هاب^۲) شرکت می‌کنند.
- ویژگی جهان کوچک^۳: پروتئین‌های یک شبکه PPI دارای میانگین طول مسیر کم و ضرایب خوش‌های بالا هستند [۲۷] که سیگنال‌های هر گره در شبکه PPI را قادر می‌سازد تا از طریق چند 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 ماذول‌های عملکردی^۴: شبکه PPI یک شبکه ماذولار و سلسله مراتبی می‌باشد. یک ماذول عملکردی در یک شبکه PPI یک مجموعه با بیشترین تعداد پروتئین که عملکرد یکسانی دارد، می‌باشد. بارزترین مشخصه ماذول عملکردی، ارتباط بین ساختار توپولوژیکی شبکه PPI و عملکرد پروتئین‌های آن است که مبنای بسیاری از روش‌های تشخیص ماذول عملکردی است [۲۸] [۲۹].

^۱ Scale-free distribution

^۲ Hub

^۳ Small-world property

^۴ Functional modular network

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^۱ یکی از شاخه‌های هوش مصنوعی^۲ 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 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.

به طور کلی، روش‌های یادگیری ماشین به سه دسته‌ی اصلی یادگیری نظارت شده^۳، یادگیری بدون نظارت^۴ و یادگیری تقویتی^۵ تقسیم می‌شوند که هر یک متناسب با نوع داده‌ها و هدف مسئله به کار گرفته می‌شوند [۳۰].

۱ Machine Learning

۲ Artificial Intelligence

۳ Supervised Learning

۴ Unsupervised Learning

۵ Reinforcement Learning

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 بهم یا لایه‌های پیچیده‌تر تشکیل می‌شوند و به دلیل 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 به صورت 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 به صورت خودکار استخراج کنند. این قابلیت باعث شده است که یادگیری عمیق در مسائل پیچیده و داده‌محور عملکرد بهتری از خود نشان دهد [۳۱].

۳.۳.۲ یادگیری نظارت شده

یادگیری نظارت شده یکی از رایج‌ترین انواع یادگیری ماشین است که در آن الگوریتم با استفاده از داده‌های ورودی به همراه برچسب^۱ متناظر با هر نمونه آموزش می‌بیند. هدف از این فرایند، ساخت مدلی است که بتواند رابطه میان داده‌ها و برچسب‌ها را آموخته و برای داده‌های جدید و دیده نشده عملکرد مناسبی داشته باشد. در این نوع یادگیری، داده‌ها معمولاً به دو مجموعه‌ی آموزشی^۲ و آزمایشی^۳ تقسیم می‌شوند؛ بدین صورت که مدل با استفاده از داده‌های آموزشی ساخته شده و سپس با داده‌های آزمایشی مورد ارزیابی قرار می‌گیرد. به بیان ساده، وجود برچسب در کنار هر داده، ویژگی اصلی یادگیری نظارت شده محسوب می‌شود [۳۲].

^۱ Label

^۲ Train dataset

^۳ Test dataset

۴.۳.۲ یادگیری بدون نظارت

یادگیری بدون نظارت نوعی از یادگیری ماشین است که در آن الگوریتم‌ها با داده‌های بدون برچسب سروکار دارند و هدف اصلی آن‌ها کشف الگوها و ساختارهای پنهان موجود در داده‌ها است. برخلاف یادگیری نظارت شده، در این روش اطلاعاتی از خروجی مطلوب در اختیار مدل قرار نمی‌گیرد و الگوریتم تلاش می‌کند بر اساس شباهت‌ها و ویژگی‌های ذاتی داده‌ها، آن‌ها را سازماندیشی کند. از شناخته شده‌ترین الگوریتم‌های یادگیری بدون نظارت می‌توان به روش‌های خوشه‌بندی اشاره کرد که با هدف قرار دادن داده‌های مشابه در خوشه‌های یکسان به کار می‌روند [۳۲].

۵.۳.۲ گراف

یک گراف از مجموعه‌ای غیر خالی از اشیا به نام راس تشکیل شده، که آن را با V نشان می‌دهیم، و مجموعه‌ای شامل یال‌ها، که راس‌ها را به هم وصل می‌کنند و با E نمایش می‌دهیم. یک چنین گرافی را با $G = (V, E)$ نشان می‌دهیم. اگر یال e دو راس v_1 و v_2 را به هم وصل کند می‌نویسیم $e = \{v_1, v_2\}$. 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 جفت اعضایی

همانند V تشکیل شده باشد و ترتیب این دو تایی‌ها نشان‌دهنده جهت یال مربوطه باشد. به این صورت برای هر یال جهت نیز درنظر 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. شبکه‌های عصبی گرافی، دسته‌ای از شبکه‌های عصبی هستند که برای مدیریت داده‌های سازماندهی شده در ساختارهای گراف طراحی شده‌اند. شبکه‌های عصبی گرافی بر پایه مکانیسم 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 بازنمایی‌های 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 می‌نمند. در مرحله بعد شبکه باید اطلاعات جمع‌آوری شده را با اطلاعات موجود گره ادغام کند و بازنمایی جدیدی از گره مورد نظر ارائه دهد. به صورت کلی این مرحله از انتقال پیام را نیز بروزرسانی^۴ می‌نمند. به طور خلاصه در یکبار انتقال پیام مراحل زیر طی می‌شوند:

$$h_u^{(k+1)} = \text{UPDATE}^{(k)}(h_u^{(k)}, \text{AGGREGATE}(\{h_v^{(k)}, \forall v \in N(u)\})) \quad (1.2)$$

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

با افزایش این مراحل، بازنمایی هر گره داده‌های بیشتری از گره‌های دورتر از خود در گراف خواهد داشت. پس از اولین تکرار ($k = 1$)، هر بازنمایی گره اطلاعات مربوز به همسایگی تک گامی خود

Graph neural networks - GNNs^۱

Message passing^۲

Aggregate^۳

Update^۴

را حفظ می‌کند، که ممکن است در گراف از طریق مسیری به طول ۱ قابل دسترسی باشد [۳۵]. بعد از دومین تکرار ($k = 2$)، بازنمایی هر گره شامل اطلاعاتی از همسایگی با دو گام است؛ به طور کلی، پس از k مرحله، بازنمایی هر گره می‌تواند شامل داده‌هایی از گره‌هایی با فاصله $hop - k$ از خود باشد. براساس مکانیزم انتقال پیام در شبکه‌های عصبی گرافی، این شبکه‌ها در تولید بازنمایی‌هایی که هم اطلاعات مربوط به ساختار گراف و هم ویژگی‌های گره‌ها را حفظ کنند بسیار موفق بوده‌اند. و 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 [۳۶].

۷.۳.۲ شبکه‌های عصبی گرافی پیچشی

پژوهش کیف و ولینگ [۳۷] با هدف ارائه مدلی ساده، مقیاس‌پذیر و قابل اجرا برای یادگیری روی گراف‌های بزرگ، شبکه‌های عصبی گرافی پیچشی را معرفی کرد. روش‌های طیفی پیشین مبتنی بر تجزیه‌ی لایه‌ای گراف بوده و نیازمند محاسبه‌ی مفادیر 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 است. تابع $\sigma(\cdot)$ نشان‌دهنده یک تابع فعال‌سازی است که برای مثال می‌تواند تابع $\text{ReLU}(\cdot) = \max(0, \cdot)$ به صورت $\text{ReLU}(\cdot)$ باشد. ماتریس

^۱Self-connections

$H^{(k)} \in \mathbb{R}^{N \times D}$ نمایش دهنده فعال سازی ها در لایه k ام بوده و $X = H^{(0)}$ به عنوان ورودی اولیه شبکه در نظر گرفته می شود. با توجه به تعریف ارائه شده در بخش شبکه های عصبی گرافی و مکانیزم انتقال پیام، توابع UPDATE و AGGREGATE، به ترتیب از معادله های $S = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2} H^{(k)}}$ و $\sigma(SW^{(k)})$ پیروی می کنند.

با وجود سادگی و کارایی بالا، شبکه های عصبی گرافی پیچشی دارای محدودیت هایی نیز هستند. از جمله این محدودیت ها می توان به پدیده هی هموار سازی بیش از حد^۱ در صورت افزایش تعداد لایه ها اشاره کرد که در آن نمایش گره ها به مرور شباهت زیادی به یکدیگر پیدا می کنند و توان تفکیک مدل کاهش می یابد. علاوه بر این، در این شبکه ها، تمامی گره های همسایه به یک اندازه در تشکیل بازنمایی جدید نقش دارند و این در حالی است که در بسیاری از گراف ها میزان اهمیت تمامی گره ها یکسان نیست. این محدودیت ها انگیزه ای برای توسعه مدل های پیشرفته تر شبکه های عصبی گرافی در پژوهش های بعدی بوده است.

۸.۳.۲ شبکه های عصبی گرافی توجه محور

ولیکویک و همکاران [۲۸]، با هدف رفع محدودیت های شبکه های عصبی گرافی پیچشی در تخصیص وزن یکسان به تمامی همسایه ها، مدل شبکه های عصبی گرافی با مکانیزم توجه^۲ را معرفی کردند. این مدل امکان یادگیری وزن های متفاوت برای هر یال همسایگی را فراهم می کند و بدین ترتیب اهمیت نسبی گره های همسایه در تشکیل بازنمایی هر گره به صورت داده محور تعیین می شود. این ویژگی باعث می شود GAT توافقی مدل سازی پیچیدگی های ساختاری گراف هایی با ارتباطات غیرهمگن را داشته باشد، در حالی که GCN تمامی همسایه ها را با وزن یکسان در نظر می گیرد.

در GAT، انتقال پیام با استفاده از مکانیسم توجه خود - توجهی^۳ در معادله ۳.۲ تعریف می شود.

Over-smoothing^۱

Graph Attention Network — GAT^۲

Self-attention^۳

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

که در آن، $h_i^{(k+1)}$ نمایش به روز شده گره i ، W ماتریس وزن قابل آموزش و σ تابع فعال سازی است. ضریب توجه α_{ij} اهمیت گره j را نسبت به گره i مشخص می کند و با استفاده از یک شبکه کوچک خطی و تابع LeakyReLU مطابق معادله ۴.۲ محاسبه می شود.

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

که در آن a بردار وزن قابل آموزش برای مکانیزم توجه و | عملگر ادغام ویژگی های گره ها است. به این ترتیب، تابع AGGREGATE در GAT به صورت جمع وزنی گره های همسایه با ضرایب توجه بوده و تابع UPDATE شامل اعمال ماتریس وزن و تابع فعال سازی روی مقدار جمع وزنی شده است.

یکی از نوآوری های مهم GAT استفاده از توجه چندسر^۱ است. در این روش، K مکانیسم توجه مستقل بر روی همان لایه اعمال می شود و نتایج هر سر یا با هم ادغام می شوند (الحاق^۲ یا میانگین گیری) تا نمایی غنی تر و پایدار تر از ویژگی های گره ها تولید شود که در معادله ۵.۲ نمایش داده شده است.

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

Multi-head attention ^۱
Concat ^۲

استفاده از مکانیزم توجه چند سر، علاوه بر افزایش ظرفیت مدل، باعث کاهش حساسیت شبکه به نویز و نوسانات محلی گراف می‌شود و کمک می‌کند تا مدل بتواند اطلاعات مفیدی از همسایگان مختلف در سطوح گوناگون استخراج کند.

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 تاثیر 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 بیش از حد 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 [۳۹] معرفی شدند. این پژوهش به یکی از چالش‌های اساسی در شبکه‌های عصبی گرافی اشاره می‌کند که ناشی از استفاده از بازنمایی‌های مبتنی بر فاصله‌های ثابت همسایگی برای تمام گره‌ها است. در چنین رویکردی، تمامی گره‌ها صرف نظر از موقعیت ساختاری خود در گراف، با تعداد یکسانی از لایه‌ها پردازش می‌شوند؛ در حالی که ساختار زیرگرافی گره‌ها می‌تواند به طور قابل توجهی با یکدیگر متفاوت باشد.

نویسندهای این پژوهش بر این باورند که به منظور استخراج بازنمایی مناسب‌تر برای هر گره، لازم است مرتبه‌ی بازنمایی آن متناسب با ویژگی‌های ساختاری همان گره انتخاب شود. به بیان دیگر، برخی گره‌ها برای دستیابی به بازنمایی 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 [۳۷] 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 تجمعی اطلاعات گستره‌تر و یادگیری بازنمایی‌های 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 [۴۰]. با این حال، در شبکه‌های عصبی گرافی، حتی با وجود استفاده از اتصال‌های باقیمانده، مسئله‌ی همگن‌شدن بیش از حد بازنمایی 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 بازنمایی 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.

mekanizm انتقال دانش بین لایه‌ها مستقل از نوع معماری بوده و قابلیت اعمال بر روی انواع شبکه‌های عصبی گرافی را دارد. با در نظر گرفتن بازنمایی‌های استخراج شده از لایه‌های مختلف شبکه برای گره

^۷ به صورت

$$h_v^{(1)}, h_v^{(2)}, \dots, h_v^{(k)},$$

این سازوکار اقدام به ترکیب این بازنمایی‌ها با استفاده از روش‌های مختلفی نظیر بیشینه‌گیری^۲، الحق^۳، یا به کارگیری شبکه‌های حافظه‌ی کوتاه‌مدت بلندمدت مبتنی بر سازوکار توجه^۴ می‌کند. بدین ترتیب،

Residual Connection ^۱

Max Pooling ^۲

Concatenation ^۳

LSTM-attention ^۴

بازنمایی 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^۱ نیمه ناظارتی^۲ برای یادگیری ویژگی‌ها از روی گراف است. این الگوریتم به طور مستقیم از الگوریتم یادگیری بازنمایی کلمات Word2Vec [۴۱] که در زمینه پردازش زبان طبیعی^۳ 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 این الگوریتم یادگیری یک بازنمایی d بعدی برای هر گره است [۴۲]. این روش در ابتدا اقدام به جایگشت تصادفی بر روی گراف به کمک الگوریتم‌های نمونه برداری اول سطح^۴ و اول عمق^۵ می‌کند. انتخاب روش مناسب نمونه برداری از اهمیت بالایی برخوردار است. در نمونه برداری اول سطح، هدف اصلی استخراج بازنمایی‌های مشابه برای گره‌هایی است که از قوانین ساختاری یکسانی پیروی می‌کنند؛ در حالی که نمونه برداری اول عمق برایجاد بازنمایی‌های 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^۶ مطابق با معادله ۶.۲ تعریف می‌شود.

$$\max_f \sum_{u \in V} \log Pr(N_S(u) | f(u)) \quad (6.2)$$

Scalable^۱

Semi-Supervised^۲

Natural Language Processing - NLP^۳

Breadth-first sampling - BFS^۴

Depth-first sampling - DFS^۵

Loss function^۶

۱۱.۳.۲ خوشه بندی در گراف های با گره های دارای ویژگی

با فرض گراف $G = (V, E, F)$ که در آن V مجموعه گره ها، E مجموعه یال ها است و F ماتریس ویژگی های گره ها می باشد، یک خوشه بندی از گراف G را می توان با C نشان داد که مجموعه ای از زیر مجموعه های V است، به صورتی که $C_i \in C$ ؛ $C_i \subset V$. هدف از خوشه بندی این است که خوشه هایی که هم از نظر ساختاری و هم از نظر ویژگی های گره ها بهم بیشترین شباهت را دارند، پیدا کنیم. همچنین خوشه های ایجاد شده باید از نظر ارتباط یال های داخل خوشه چگال و در ارتباط یال ها با دیگر خوشه ها تنک باشند. نکته مهم دیگر در این قسمت وجود و یا عدم وجود همپوشانی در بین خوشه ها می باشد که به خوشه بندی بدون همپوشانی، افزایش بندی^۱ نیز می گویند. به عبارت دیگر در افزایش بندی 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 خوشه بندی با همپوشانی چنین شرطی الزامی نیست.

۱۲.۳.۲ دسته بندی و روش های کلی خوشه بندی گراف

روش های خوشه بندی گراف را می توان از دیدگاه های مختلفی تقسیم بندی کرد. این تقسیم بندی ها بر اساس معیارها و ویژگی های خاصی صورت می گیرند که به نحوه برخورد با داده های گرافی، نوع اطلاعات استفاده شده، و تکنیک های به کار گرفته شده بستگی دارد. در این پژوهش از آنجایی که نوع گراف ورودی مشخص است و قصد خوشه بندی گراف های PPI با گره های دارای ویژگی را داریم، روش های خوشه بندی را بر اساس روش مورد استفاده تقسیم بندی می کنیم:

- روش های طیفی^۲ : از مقادیر ویژه^۳ ماتریس لاپلاسین یا مجاورت برای یافتن خوشه ها استفاده

Partitioning^۱
Spectral clustering^۲
Eigenvalues^۳

می‌کنند.

- روش‌های فاکتورگیری ماتریسی^۱ : از روش‌های تجزیه ماتریسی مانند تجزیه نامنفی ماتریس^۲ یا تجزیه مقدار تکین^۳ برای ایجاد امبدینگ و خوشه‌بندی استفاده می‌کنند.
- روش‌های سلسله‌مراتبی^۴ : گراف را به صورت سلسله مراتبی خوشه‌بندی می‌کنند که به دو روش تقسیمی و تجمعی دسته‌بندی می‌شوند.
- روش‌های مبتنی بر امبدینگ^۵ : ابتدا گره‌ها به فضای برداری کم بعد نگاشت می‌شوند و سپس خوشه‌بندی روی این فضای برداری انجام می‌شود و تمرکز اصلی در این روش‌ها یافتن بازنمایی مناسب برای خوشه‌بندی گراف است. (Node2Vec, DeepWalk, GCN, GNN).
- روش‌های بدون امبدینگ^۶ : مستقیماً از ساختار گراف برای خوشه‌بندی استفاده می‌شود بدون اینکه گره‌ها به فضای برداری منتقل شوند (Louvain, graph - cut based).

۴.۲ معیارهای ارزیابی

در این قسمت به بررسی معیارهای ارزیابی عملکرد الگوریتم‌های شناسایی مجموعه‌های پروتئینی می‌پردازیم. در بین معیارهای موجود، معیارهای دقیق^۷، بازیابی^۸، صحبت^۹، 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

Matrix factorization	^۱
Non-negative matrix factorization	^۲
Singular value factorization	^۳
Hierarchical clustering	^۴
Embedding-based methods	^۵
Non-embedding methods	^۶
Precision	^۷
Recall	^۸
Accuracy	^۹

نیاز است، می‌پردازیم:

۱.۴.۲ شباهت همسایگی^۱

با در نظر گرفتن P به عنوان مجموعه‌ای از مجموعه‌های پروتئینی شناسایی شده توسط الگوریتم، عملکرد الگوریتم به وسیله تعداد مجموعه‌های پروتئینی مشترک بین P و مجموعه‌ای از مجموعه پروتئین‌های مرجع^۲ 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:

$$NA(p, b) = \frac{|V_p \cap V_b|^{\omega}}{|V_p| \times |V_b|} \quad (7.2)$$

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^۳ از قبل تعیین شده (معمولًا ۰/۲۵) در نظر گرفته می‌شود که شباهت همسایگی‌های بالاتر از آستانه به معنی یکسانی دو مجموعه است. همچنانی 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.

$$N_{cp} = \{p | p \in P, \exists b \in B, NA(p, b) \geq \omega\} \quad (8.2)$$

Neighborhood affinity^۱
Reference protein complex^۲
Threshold^۳

$$N_{cb} = \{b | b \in B, \exists p \in P, NA(p, b) \geq \omega\} \quad (9.2)$$

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.

$$Precision = \frac{N_{cp}}{|P|} \quad (10.2)$$

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.

$$Recall = \frac{N_{cb}}{|B|} \quad (11.2)$$

۴.۴.۲ امتیاز F

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

¹ Harmonic mean

$$F-score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (12.2)$$

۵.۴.۲ صحت

معیار صحت به کمک دو معیار دیگر حساسیت خوشبندی^۱ و ارزش پیش‌بینی مثبت خوشبندی^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 j ام یافت می‌شوند و همچنین N به عنوان تعداد پروتئین‌های مجموعه پروتئینی مرجع i ، می‌توانیم Sn و PPV را به صورت مقابل تعریف کنیم:

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

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

از نظر مفهومی، معیار PPV نشان‌دهنده نسبت مجموع بیشینه پروتئین‌های تطبیق‌یافته هر مجموعه پروتئینی پیش‌بینی شده با مجموعه‌های پروتئینی مرجع، به تعداد کل پروتئین‌های تطبیق‌یافته در مجموعه‌های پروتئینی پیش‌بینی شده است. از سوی دیگر، معیار Sn بیان‌کننده 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:

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

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 گراف‌های دارای گره ویژگی

پژوهش وحید جان‌ثاری و همکارانش [۴۳]، یک الگوریتم بر پایه تجزیه نامنفی ماتریسی^۲ به منظور خوشه‌بندی گراف‌های ویژگی‌دار معرفی می‌کند. روش آن‌ها ابتدا اطلاعات ساختاری که توسط ماتریس

Non-negative matrix factorization ^۲

همسایگی^۱ نشان داده می شود را به کمک تجزیه نامنفی متقارن ماتریس^۲ و اطلاعات ویژگی های گره ها را به کمک تجزیه نامنفی بازتابی ماتریس^۳ به یک فضای کم بعد مختص خوش بندی (هم بعد با تعداد خوش بندی) به صورت جداگانه انتقال می دهد که درجه عضویت هر گره به هر خوش بندی را نمایش می دهد. همینطور به منظور حفظ ثبات در خوش بندی در هر دو فضای اقدام به نزدیک کردن این دو ماتریس به کمک تابع هدف می کند که به صورت مقابل تعریف شده است:

$$J_{of} = \min \left\| A - VV^T \right\|_F^2 + \alpha \left\| VV^T - UU^T \right\|_F^2 + \left\| F - UU^T F \right\|_F^2 \quad (1.3)$$

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

که در تابع هدف، A ماتریس همسایگی، $V \in R^{n \times k}$ ماتریس حاصل از تجزیه نامنفی متقارن ماتریس است. همینطور با در نظر گرفتن M (ماتریس شباهت^۴ گره ها براساس ماتریس ویژگی ها) به صورت $A = UU^T; U \in R^{n \times k}$ و عبارت سوم در بهینه سازی که به صورت مقابل بیان شده است: $|F - UU^T F|$ در واقع اقدام به استفاده از ویژگی خودبیانگری^۵ داده ها کرده اند، که در نتیجه روش بیان شده را MF می توان یک روش ترکیبی از خوش بندی زیر فضا^۶ و تجزیه نامنفی ماتریس در نظر گرفت.

در پژوهشی دیگر توسط کانگ و همکارانش [۴۴]، یک روش بر پایه شبکه های پیچشی گرافی^۷ و خوش بندی 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 ویژگی های گره ها و ساختار گراف به ماتریس

Adjacency matrix	^۱
Symmetric non-negative matrix factorization	^۲
Projective non-negative matrix factorization	^۳
Similarity matrix	^۴
Self-expression	^۵
Subspace clustering	^۶
Graph convolutional networks	^۷
Low-pass filter	^۸

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

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

همچنین در فرمول بالا k یک هایپر پارامتر است که میزان مرتبه مجاورت بازنمایی به دست آمده را مشخص می‌کند به عبارت دیگر مقادیر کوچک‌تر k دید محلی تری به ساختار گراف دارند و بالعکس. L که L ماتریس لاپلاسی نرمال شده¹ می‌باشد. در مرحله بعد برای اعمال خوشبندی 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 خوشبندی طیفی روی آن اعمال می‌شود. یکی از مشکلات این روش انتخاب مناسب هایپر پارامتر K است که به طور مستقیم برخروجی الگوریتم تاثیر می‌گذارد که توسط پژوهش دیگری که توسط ژانگ و همکارانش [۴۵] انجام شده است، دو استراتژی AGC و IAGC برای پیدا کردن مقدار مناسب k ارائه شده است.

یکی از مشکلات روش‌های بر پایه بازنمایی این است که دو فرآیند بازنمایی‌ها داده‌ها و خوشبندی از یکدیگر مستقل‌اند در نتیجه نمی‌توان اطمینان داشت که بازنمایی‌های ایجاد شده برای وظیفه موردنظر (در اینجا خوشبندی) مناسب هستند و همچنین نمی‌توان الگوریتم بازنمایی را بر اساس خطای خوشبندی به طور مناسب به روزرسانی نمود. از این روی، وانگ و همکاران [۴۶] یک روش خوش

Normalized Laplacian matrix ۱

بندی یکپارچه توجه محور بر پایه شبکه عصبی گراف ارائه داده‌اند که مرحله بازنمایی و خوشبندی 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 بازنمایی 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 بازنمایی خود بهینه‌ساز است که در آن به طور مکرر نقاط مربوط به هر خوش برا اساس مقدار اطمینان تعلق به خوش بروزرسانی می‌شود و به طور همزمان بازنمایی‌ها را نیز به وسیله آن اصلاح می‌کند.

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. به طور کلی الگوریتم‌های پیش‌بینی مجموعه‌های پروتئینی را می‌توان به دو دسته تقسیم کرد:

روش‌های بر پایه شبکه^۴:

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

- روش‌های تقسیمی^۵: این دسته از روش‌ها، شبکه را به زیر شبکه‌ها تقسیم می‌کنند و این عمل را تا رسیدن به درجه دلخواه خوش بندی تکرار می‌کنند. معروف‌ترین الگوریتم این دسته الگوریتم

Graph attentional^۱

Decoder^۲

Encoder^۳

Network-based methods^۴

Divisive methods^۵

خوشبندی مارکوف^۱ [۴۷] است که زیر شبکه‌ها را به کمک قدم تصادفی^۲ در شبکه پیدا می‌کند.

- روش‌های تجمعی^۳: با مجموعه کوچکی از پروتئین‌ها شروع کرده و با ترکیب آن‌ها اقدام به پیدا کردن مجموعه‌های پروتئینی نهایی می‌کند. الگوریتم CPNM [۴۸] یکی از الگوریتم‌های این دسته است که از امبینگ موتیف‌های^۴ شبکه به منظور پیدا کردن نقش پروتئین‌ها استفاده 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 [۴۹] است. این الگوریتم ابتدا پروتئین‌های با درجه بالاتر را به عنوان پروتئین‌های هسته^۵ (پروتئین‌های آغازین) در نظر گرفته می‌گیرد. سپس زیرگروه‌هایی از گره‌ها با بیشترین انسجام برای گره‌های هسته انتخاب می‌شوند. در انتهای 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^۶ [۵۰] است که در سه مرحله انجام می‌شود. این الگوریتم ابتدا گره‌ها را وزن دهی می‌کند، سپس به شناسایی مجموعه‌ها می‌پردازد و در انتهای نیز اقدام به اضافه / حذف کردن پروتئین‌ها به/از مجموعه‌های شناسایی شده با توجه به یک معیار اتصال می‌کند.

روش‌های مبتنی بر آگاهی از زمینه‌های زیستی^۷:

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 [۵۱]، الگوریتم

^۱ Markov clustering algorithm

^۲ Random walk

^۳ Agglomerative methods

^۴ Motif

^۵ Seed

^۶ detection complex olecularM

^۷ Biological-context-aware-based methods

PCIA را توسعه داده‌اند که از ترکیب اطلاعات GO در کنار ساختار شبکه استفاده می‌کند. پژوهش دیگر ژانگ و همکارانش [۵۲] رابطه‌ی بین شکل گیری مجموعه‌های پروتئینی و هم بیانی پروتئین‌ها را نشان داده است.

- روش‌های هسته-اتصال^۱: 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. الگوریتم COACH [۵۳] یکی از شناخته شده‌ترین الگوریتم‌های این دسته است که از دو مرحله شناسایی پروتئین‌های هسته‌ای و اضافه کردن پروتئین‌ها به پروتئین‌های هسته‌ای تشکیل شده است. تمرکز این الگوریتم بر ایجاد مجموعه‌های پروتئینی است که از نظر زیستی نیز با معنی باشند. الگوریتم CORE [۵۴] نیز از سه مرحله، پیش‌بینی پروتئین‌های هسته‌ای، حذف هسته‌های با اهمیت پایین (بر اساس یک معیار اتصال)، و محاسبه اهمیت مجموعه‌های شناسایی شده، تشکیل شده است. اخیراً نیز الگوریتم CO-DPC از این دسته بنده ارائه شده است که از نمایه بیان ثنی در کنار شبکه PPI استفاده می‌کند.
- الگوریتم‌های مبتنی بر اطلاعات عملکردی^۲: دسته دوم الگوریتم‌ها روش‌های مبتنی بر اطلاعات عملکردی هستند که از اطلاعات ناهمگون پروتئین‌ها به منظور شناسایی مجموعه‌های با معنی استفاده می‌کنند. یکی از الگوریتم‌های این دسته، الگوریتم PCP [۵۵] است که از اطلاعات ساختاری به منظور وزن‌دهی شبکه PPI استفاده می‌کند. سپس ابتدا اقدام به شناسایی کلیک‌های بیشینه^۳ در شبکه PPI کرده، در مرحله بعد چگالی بین خوشه‌ها را محاسبه می‌کند و در نهایت اقدام به ترکیب جزئی کلیک‌ها می‌کند.

لازم به ذکر 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

Core-attachment^۱

Functional-information-based^۲

Maximal clique^۳

امبینگ [۵۶] و تجزیه ماتریسی [۵۷] اشاره کرد که به منظور شناسایی مجموعه‌های پروتئینی نیز مورد استفاده قرار گرفته‌اند.

فصل ۴

روش

در این فصل به بررسی روش پیشنهادی بر پایه شبکه‌های عصبی گرافی به منظور خوشه‌بندی شبکه تعاملات پروتئین-پروتئین می‌پردازیم. در ابتدای این فصل نگاهی به مجموعه داده‌های موجود می‌کنیم و دلیل انتخاب آن‌ها را برای آزمایش روش پیشنهادی بررسی می‌کنیم.

۱.۴ مجموعه داده

در دهه گذشته، داده‌های PPI از طریق روش‌های آزمایشگاهی با خروجی بالا^۲ مانند سیستم‌های دوگانه هیبریدی^۳ [۵۹]، طیف‌سنجی جرمی^۴ [۶۰] به شدت غنی شده‌اند. همچنین، روش‌های متن کاوی^۵ برای ایجاد شبکه‌های PPI نیز به صورت گسترده استفاده شده‌اند [۸] [۶۱] [۶۲]. به طور کلی می‌توان منابع داده PPI را به دسته‌های آزمایشگاهی، پایگاه داده‌های ایجاد شده به کمک روش‌های محاسباتی و همچنین پایگاه داده‌های ادغام شده تقسیم بندی کرد. به عنوان مثال می‌توان به برخی از این مجموعه

۲ High-throughput

۳ Two-hybrid systems

۴ Mass spectrometry

۵ Text mining

داده‌های تعامل پروتئین پروتئین مانند Biogrid [۶۳] DIP [۶۴] و MIPS [۶۵] Collins [۶۶] اشاره کرد. برای صحت سنجی از مجموعه‌های پروتئینی یافت شده نیز از مجموعه داده‌هایی شامل مجموعه‌های پروتئینی شناخته شده مانند CYC2008 و یا MIPS می‌توان استفاده کرد.

۲.۴ روش پیشنهادی

روش پیشنهادی ما در این پژوهش به کمک استفاده از شبکه‌های عصبی گرافی یک بازنمایی مناسب به منظور خوشه‌بندی شبکه تعاملات پروتئین-پروتئین با در نظر گرفتن ویژگی‌های زیستی و بیان ژنی ایجاد می‌کند که هیمنظور امکان خوشه‌بندی همپوشان را نیز ممکن می‌سازد. روش پیشنهادی از مدل مولد احتمالی برنولی پواسون کمک می‌گیرد و تابع هزینه جدیدی را بر این پایه معرفی می‌کنیم. روش پیشنهادی شامل سه مرحله است که در ادامه به بررسی بیشتر این مراحل می‌پردازیم.

۱.۲.۴ مرحله اول: استفاده از شبکه عصبی گرافی به منظور ایجاد ماتریس وابستگی

با فرض داشتن گراف نود ویژگی دار G که می‌توان آن را با دو ماتریس مجاورت A و ویژگی نودهای X نمایش داد، یک شبکه عصبی گرافی کانولوشنی دو لایه به منظور ایجاد ماتریس وابستگی F در نظر می‌گیریم:

$$F = GNN_{\theta}(A, X) \quad (1.4)$$

$$\tilde{A} = A + I_N \quad (2.4)$$

$$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij} \quad (3.4)$$

$$\hat{A} = \tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} \quad (4.4)$$

$$F = ReLU(\hat{A} ReLU(\hat{A} X W^{(1)}) W^{(2)}) \quad (5.4)$$

مدل شبکه عصبی گرافی پیشنهادی دو تفاوت با شبکه‌های عادی دارد:

- استفاده از لایه نرم‌السازی دسته‌ای بعد از لایه اول گراف کانولوشن

- اعمال L_2 regularization بر روی ماتریس وزن‌ها ($W^{(1)}$ و $W^{(2)}$)

۲.۲.۴ مرحله دوم: بهینه‌سازی وزن‌های شبکه عصبی گرافی

در ابتدا باید نگاهی به مفهوم مدل مولد برنولی پواسون داشته باشیم، این مدل سعی بر بازسازی گراف به کمک ماتریس وابستگی F به صورت مقابل دارد:

$$A_{uv} \sim \text{Bernoulli}(1 - e^{-F_u F_v^T}) \quad (6.4)$$

حال می‌توان با استفاده از مدل برنولی پواسون به محاسبه likelihood $p(A|F)$ یا با فرمولاسیون مقابل عمل کنیم:

$$P(A|F) = \prod_{A_{uv} \in E} 1 - e^{-F_u F_v^T} \times \prod_{A_{uv} \notin E} e^{-F_u F_v^T} \quad (7.4)$$

در مرحله بعد به منظور ایجاد تابع هزینه، اقدام به اعمال \log - می‌کنیم. در نتیجه به فرمول مقابل می‌رسیم:

$$-\log p(A|F) = - \sum_{A_{uv} \in E} \log(1 - \exp(-F_u F_v^T)) + \sum_{A_{uv} \notin E} F_u F_v^T \quad (8.4)$$

حال می‌توانیم ادعا کنیم که تابع هزینه ما با $\log(A|F)$ – برابر می‌کند.

$$L(F) = - \sum_{A_{uv} \in E} \log(1 - \exp(-F_u F_v^T)) + \sum_{A_{uv} \in E} F_u F_v^T \quad (9.4)$$

تابع هزینه این است که ماتریس همسایگی در بیشتر موارد یک ماتریس به شدت تنک می‌باشد. از این روی مقدار عبارت دوم در رابطه بیشتر از قسمت اول می‌شود. به همین دلیل اقدام به استفاده از مقدار امید ریاضی هر یک از عبارات با توزیع یکنواخت بر روی تمامی یال‌ها می‌کنیم.

$$L(F) = -E_{(U,V) \sim P_E} [\log(1 - \exp(-F_u F_v^T))] + E_{(u,v) \sim P_N} [F_u F_v^T] \quad (10.4)$$

که در آن P_E توزیع یکنواخت بر روی یال‌ها و P_N یک توزیع یکنواخت بر روی دو راسی است که بین آن‌ها یال وجود ندارد. در نهایت می‌توان تابع هزینه حاصل را به صورت مقابل نمایش داد:

$$\theta^* = \operatorname{argmin}_{\theta} L(GNN(A, X)) + \lambda_1 \|W^{(1)}\|_2 + \lambda_2 \|W^{(2)}\|_2 \quad (11.4)$$

۳.۲.۴ مرحله سوم: تخصیص نودها به خوشه‌ها

در نهایت با پیدا کردن پارامترهای مدل، اقدام به پیش‌بینی ماتریس وابستگی F می‌کنیم و برای تخصیص نودها به خوشه‌ها یک آستانه φ در نظر می‌گیریم:

$$F_{uc} = \begin{cases} 1 & \text{if } F_{uc} > \varphi \\ 0 & \text{otherwise} \end{cases} \quad (12.4)$$

كتاب نامه

- [1] V, Manila M. A literature survey on bioinformatics. *IJIREEICE International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, February 2023.
- [2] Ji, Junzhong, Zhang, Aidong, Liu, Chunian, 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.
- [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, Goliae, 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. Bulkrnabert: 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] Gene Ontology overview, December 2025. [Online; accessed 17. Dec. 2025].
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] Wagner, Günter P, Pavlicev, Mihaela, and Cheverud, James M. The road to modularity. *Nature Reviews Genetics*, 8(12):921–931, 2007.
- [30] 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.
- [31] Janiesch, Christian, Zschech, Patrick, and Heinrich, Kai. Machine learning and deep learning. *Electronic markets*, 31(3):685–695, 2021.
- [32] Dridi, Salim. Supervised learning-a systematic literature review. *preprint*, Dec, 2021.

[33] 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, Moltakeran, .*Mathematics Discrete in Topics* Esmail. Babalian, [۳۴]

.۲۰۰۷

[35] Rong, Yu, Huang, Wenbing, Xu, Tingyang, and Huang, Junzhou. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*, 2019.

[36] 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.

[37] Kipf, TN. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[38] Veličković, Petar, Cucurull, Guillem, Casanova, Arantxa, Romero, Adriana, Lio, Pietro, and Bengio, Yoshua. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[39] 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.

- [40] 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.
- [41] Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [42] 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.
- [43] 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.
- [44] 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.
- [45] 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.
- [46] 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.

- [47] 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.
- [48] Patra, Sabyasachi and Mohapatra, Anjali. Protein complex prediction in interaction network based on network motif. *Computational Biology and Chemistry*, 89:107399, 2020.
- [49] Nepusz, Tamás, Yu, Haiyuan, and Paccanaro, Alberto. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*, 9(5):471–472, 2012.
- [50] 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.
- [51] 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.
- [52] 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.

- [53] 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.
- [54] 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.
- [55] 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.
- [56] 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.
- [57] 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.
- [58] 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.

- [59] 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.
- [60] 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.
- [61] 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.
- [62] 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.
- [63] 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.
- [64] 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.
- [65] Collins, Sean R, Kemmeren, Patrick, Zhao, Xue-Chu, Greenblatt, Jack F,

$$\omega^\circ$$

- 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.
- [66] 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.

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*



**Institute for Advanced Studies
in Basic Sciences**
Gava Zang, Zanjan, Iran

**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

December 19, 2025