وزارت علوم، تحقیقات و فناوری
دانشگاه تحصیلات تکمیلی علوم پایه
گاوازنگ - زنجان

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**

۲۷ آذر ۱۴۰۴


بسم الله الرحمن الرحیم

تقدیم به آنهایی که می خواهند بیشتر بدانند.

تشکر 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 ماژول‌های عملکردی، خوشه‌بندی گراف‌های دارای ویژگی*

فهرست مطالب

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

بیوانفورماتیک 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 سزایی 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 مجموعه‌ای از پروتئین‌ها انجام می‌گیرد که ماژول‌های عملکردی نامیده می‌شوند. شناخت هر چه بهتر این مجموعه‌های پروتئینی به درک بهتر ما از فرایندهای زیستی و همچین 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.

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 دی‌اِن‌اِی[۳]، آر‌اِن‌اِی[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 پروژه توالی‌یابی ژنوم انسان و ورود به دوره‌ی پساژنی[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 ـ 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].

---

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

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 زیرشبکه‌های به هم متصل با تراکم بالا از پروتئین‌ها می‌تواند به عنوان ماژول‌های 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 مجموعه های پروتئینی مختلف را بررسی کرد و یا حتی مجموعه‌های پروتئینی ناشناخته‌ای را کشف کرد [۲].

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 پساژنی، ارائه الگوریتم‌های بهینه به منظور شناسایی مؤثر ماژول‌های 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 اضاقه 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 ژنی هستند برخی از پژوهش‌ها از بیان ژنی 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 خوشه‌بندی گرافی داریم که به طور مناسب ویژگی‌های پروتئین‌ها را در کنار ساختار شبکه برای خوشه‌بندی مجموعه‌های پروتئینی در نظر گیرد، به این دسته از الگوریتم‌ها، خوشه‌بندی گراف‌های دارای ویژگی[۱] می‌گویند [۱۰]. آز آنجایی که یک پروتئین می‌تواند در چندین فرآیند زیستی نقش داشته باشد، در نتیجه بسیاری از مجموعه‌های پروتئینی شامل پروتئین‌های 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 ماژول‌های 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 صرفا 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 ماژول‌های عملکردی هم پوشان با دید گرافی به شبکه‌های PPI در کنار استفاده از داده‌های تکمیلی بانک‌های داده‌های زیستی، با روش خوشه‌بندی گراف با گره‌های ویژگی‌دار را داریم. در ادامه اهمیت موضوع، پرسش‌هایی که در این پژوهش به دنبال جواب آن‌ها هستیم و همچنین روش پیشنهادی را به صورت

---

[۱] Attributed graph clustering

[۲] End-to-end

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 دی‌اِن‌اِی 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 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].

۴.۱ پرسش‌های پژوهش

بعد از تکمیل قسمت روش

[۱] Pathway

- الگوریتم‌های موجود به‌ویژه 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 ماژول‌های تعاملات پروتئینی اضافه کرد؟
- ترکیب اطلاعات توپولوژی و اطلاعات زیستی تا چه حد کشف ماژول‌های پروتئینی را بهبود می‌دهد؟
- وجود همپوشانی بین ماژول‌ها در افزایش دقت در کشف ماژول‌های پروتئینی تا چه حد اثرگذار است؟

۵.۱ روش 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 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 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 به‌کاررفته 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 به‌منظور 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 فرایندهایی 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 فرایند ترجمه از روی آرای(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 به‌گونه‌ای است که تقریباً تمامی فرایندهای 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].

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 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 دی‌اِن‌اِی، رونوشت آر‌اِن‌اِی پیام‌رسان و تنظیم چرخه‌ی سلولی اشاره کرد. در هر یک از این فرآیندها، گروه مشخصی از پروتئین‌ها به‌صورت 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 بیوانفورماتیک تبدیل شده است. شناسایی دقیق این ماژول‌ها کاربردهای گسترده‌ای از جمله پیش‌بینی عملکرد پروتئین‌های ناشناخته [۱۹]، درک مکانیسم‌های مولکولی بیماری‌ها

[۱] Functional Module

[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[۱]

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

۴.۲.۲ پایگاه داده هستی شناسی 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 فرایندهای 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 درحال تغییر است [۲۳].

---

۱ Gene expression

۲ Gene Ontology

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 به‌ترتیب بیانگر رابطهٔ «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 مراتبی است که شامل شبکه‌های بدون دور می‌شود و ویژگی‌ها به این صورت مرتب شده‌اند [۸].

---

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

۵.۲.۲ شبکه‌های PPI و ویژگی‌های آنها

یک شبکه PPI معمولا به صورت یک گراف بدون جهت $G = (V, E)$ نشان داده می شود که V و E به ترتیب نمایانگر پروتئین ها و تعاملات بین آنها می باشند. وزن های روی یال ها را می توان برای توصیف ویژگی های شبکه PPI، مانند ویژگی‌های توپولوژیکی یا عملکردی استفاده کرد. شبکه های PPI سه ویژگی توپولوژیکی زیر را دارند:

- توزیع بدون مقیاس[۱] : $P(k)$ مفهوم توزیع درجه یعنی احتمال اینکه یک گره در یک شبکه دقیقا k پیوند داشته باشد را نشان می دهد. یک شبکه PPI دارای توزیع درجه توانی $P(k) \sim k^{-\lambda}$ می باشد [۲۶] . این ویژگی به این معنی است که پروتئین‌های تعامل‌دار در شبکه‌های PPI به طور یکنواخت توزیع نمی شوند، 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 ماژول‌های عملکردی[۴] : 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 [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 ارائه‌شده ضروری است. بدین منظور، در ادامه مروری اجمالی بر مفاهیم و تعاریف اصلی ارائه می‌شود تا چارچوب محاسباتی پژوهش به‌صورت منسجم و شفاف تبیین گردد.

۱.۳.۲ یادگیری ماشین

۲.۳.۲ یادگیری عمیق

۳.۳.۲ یادگیری نظارت شده

۴.۳.۲ یادگیری بدون نظارت

۵.۳.۲ گراف

یک گراف از مجموعه‌ای غیر خالی از اشیا به نام رأس تشکیل شده، که آن را با V نشان می‌دهیم، و مجموعه‌ای شامل یال‌ها، که رأس‌ها را به هم وصل می‌کنند و با E نمایش می‌دهیم. یک چنین گرافی را با $G = (V, E)$ نشان می‌دهیم. اگر یال e دو رأس $v_۱$ و $v_۲$ را به هم وصل کند می‌نویسیم $e = \{v_۱, v_۲\}$ [۳۰]. 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 $(u, v);\ u, v \in V$ تشکیل شده باشد و ترتیب این دوتایی‌ها 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] می‌نامند. 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] می‌نامند. به طورخلاصه در یکبار انتقال پیام مراحل زیر طی می‌شوند:

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

---

۱ Graph neural networks - GNNs
۲ Message passing
۳ Aggregate
۴ Update

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 بازنمایی گره u در مرحله kام است.

با افزایش این مراحل، بازنمایی هر گره داده‌های بیشتری از گره‌های دورتر از خود در گراف خواهد داشت. پس از اولین تکرار ($k = ۱$)، هر بازنمایی گره اطلاعات مربوز به همسایگی تک گامی خود را حفظ می‌کند، که ممکن است در گراف از طریق مسیری به طول ۱ قابل دسترسی باشد [۳۱]. بعد از دومین تکرار ($k = ۲$)، 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 $k - hop$ 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 [۳۲].

۷.۳.۲ شبکه‌های عصبی گرافی پیچشی

پژوهش کیف و ولینگ [۳۳] با هدف ارائه مدلی ساده، مقیاس‌پذیر و قابل‌اجرا برای یادگیری روی گراف‌های بزرگ، شبکه‌های عصبی گرافی پیچشی را معرفی کرد. روش‌های طیفی پیشین مبتنی بر تجزیه‌ی لاپلاسین 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.

$$H^{(k+۱)} = \sigma(\tilde{D}^{-\frac{۱}{۲}}\tilde{A}\tilde{D}^{-\frac{۱}{۲}}H^k W^k) \qquad (۲.۲)$$

در اینجا، $\tilde{A} = A + I_N$ ماتریس مجاورت گراف بدون‌جهت 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 قابل‌آموزش مخصوص لایه kام است. تابع $\sigma(\cdot)$ نشان‌دهنده یک تابع فعال‌سازی است که برای مثال می‌تواند تابع ReLU به‌صورت $\mathrm{ReLU}(\cdot) = \max(0, \cdot)$ باشد. ماتریس $H^{(k)} \in \mathbb{R}^{N \times D}$ نمایش‌دهنده فعال‌سازی‌ها در لایه kام بوده و $H^{(0)} = X$ به‌عنوان ورودی اولیه شبکه در نظر گرفته می‌شود. با توجه به تعریف ارائه شده در بخش شبکه‌های عصبی گرافی و مکانیزم 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[۲] در صورت افزایش تعداد لایه‌ها اشاره کرد که در آن نمایش 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 بازنمایی 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 [۳۴]، با هدف رفع محدودیت‌های شبکه‌های عصبی گرافی پیچشی در تخصیص وزن یکسان به تمامی همسایه‌ها، مدل شبکه‌های عصبی گرافی با مکانیزم توجه[۳] را معرفی کردند. این مدل امکان یادگیری وزن‌های متفاوت برای هر یال همسایگی را فراهم می‌کند و بدین ترتیب اهمیت

---

۱ Self-connections

۲ Over-smoothing

۳ Graph Attention Network — GAT

نسبی گره‌های همسایه در تشکیل بازنمایی 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 مکانیسم توجه خود-توجهی[۱] در معادله ۳.۲ تعریف می‌شود.

$$h_i^{(k+1)} = \sigma\Big(\sum_{j\in\mathcal{N}(i)} \alpha_{ij} W h_j^{(k)}\Big) \tag{۳.۲}$$

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.

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(a^T[Wh_i^{(k)}|Wh_j^{(k)}]))}{\sum_{n\in\mathcal{N}(i)} \exp(\text{LeakyReLU}(a^T[Wh_i^{(k)}|Wh_n^{(k)}]))} \tag{۴.۲}$$

که در آن a بردار وزن قابل‌آموزش برای مکانیزم 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[۲] است. در این روش، K مکانیسم توجه مستقل بر روی همان لایه اعمال می‌شود و نتایج هر سر یا با هم ادغام می‌شوند (الحاق[۳] یا میانگین‌گیری) تا

[۱] Self-attention
[۲] Multi-head attention
[۳] Concat

نمایی غنی‌تر و پایدارتر از ویژگی‌های گره‌ها تولید شود که در معادله ۵.۲ نمایش داده شده است.

$$h_i^{(k+1)} = \Big\|_{m=1}^{M} \sigma\Big(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(m)} W^{(m)} h_j^{(k)} \Big) \tag{۵.۲}$$

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 قابل‌توجه، 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 به‌خوبی 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 پرشی

۱۰.۳.۲ تعبیه گره‌ها به روش Node2Vec

روش Node2Vec یک الگوریتم مقیاس پذیر[۱] نیمه نظارتی[۲] برای یادگیری ویژگی‌ها از روی گراف است. این الگوریتم به طور مستقیم از الگوریتم یادگیری بازنمایی کلمات Word2Vec [۳۵] که در

[۱] Scalable
[۲] Semi-Supervised

زمینه پردازش زبان طبیعی[۱] 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 [۳۶]. این روش در ابتدا اقدام به جایگشت تصادفی بر روی گراف به کمک الگوریتم‌های نمونه برداری اول سطح[۲] و اول عمق[۳] می‌کند. انتخاب روش مناسب نمونه‌برداری از اهمیت بالایی برخوردار است. در نمونه‌برداری اول سطح، هدف اصلی استخراج بازنمایی‌های مشابه برای گره‌هایی است که از قوانین ساختاری یکسانی پیروی می‌کنند؛ در حالی‌که نمونه‌برداری اول عمق بر ایجاد بازنمایی‌های مشابه برای گره‌هایی تمرکز دارد که به‌صورت چگال به یکدیگر متصل هستند. در عمل، بهترین راهکار استفاده از یک روش ترکیبی است؛ به‌گونه‌ای که بخشی از توالی‌ها با استفاده از نمونه‌برداری اول عمق و بخش دیگری با استفاده از نمونه‌برداری 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[۴] مطابق با معادله ۶.۲ تعریف می‌شود.

$$\max_{f} \sum_{u \in V} \log Pr(N_S(u) \mid f(u)) \tag{۶.۲}$$

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

[۱] Natural Language Processing - NLP
[۲] Breadth-first sampling - BFS
[۳] Depth-first sampling - DFS
[۴] Loss function

که خوشه هایی که هم از نظر ساختاری و هم از نظر ویژگی های 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 بین خوشه‌ها می‌باشد که به خوشه بندی بدون همپوشانی، افرازبندی[۱] نیز می‌گویند. به عبارت دیگر در افرازبندی شرط:

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

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 لاپلاسین 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 امبدینگ و خوشه‌بندی استفاده می‌کنند.

---

۱ Partitioning
۲ Spectral clustering
۳ Eigenvalues
۴ Matrix factorization
۵ Non-negative matrix factorization
۶ Singular value factorization

- روش‌های سلسله‌مراتبی[۱] : گراف را به صورت سلسله مراتبی خوشه‌بندی می‌کنند که به دو روش تقسیمی و تجمعی دسته‌بندی می‌شوند.
- روش‌های مبتنی بر امبدینگ[۲] : ابتدا گره‌ها به فضای برداری کم‌بعد 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[۳] : 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 دقت[۴]، بازیابی[۵]، صحت[۶]، 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 Hierarchical clustering
۲ Embedding-based methods
۳ Non-embedding methods
۴ Precision
۵ Recall
۶ Accuracy

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:

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

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[۳] از قبل تعیین شده (معمولا 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 مجموعه های مرجعی که حداقل با یکی از مجموعه‌های شناسایی شده الگوریتمی یکسان در نظر گرفته می‌شوند را با N_{cb} نمایش می‌دهیم [۴].

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

---

۱ Neighborhood affinity

۲ Reference protein complex

۳ Threshold

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

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:

[۱] Harmonic mean

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

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 Sn و PPV را به صورت مقابل تعریف کنیم:

$$PPV = \frac{\sum_{j=۱}^{|P|} max_{i=۱}^{|B|} |T_{ij}|}{\sum_{j=۱}^{|P|} \sum_{i=۱}^{|B|} T_{ij}}$$

$$Sn = \frac{\sum_{i=۱}^{|B|} max_{j=۱}^{|P|} |T_{ij}|}{\sum_{i=۱}^{|B|} N_i}$$

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

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

---

[۱] Clustering-wise sensitivity (Sn)

[۲] 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 جان‌نثاری و همکارانش [۳۷] ، یک الگوریتم بر پایه تجزیه نامنفی ماتریسی[۲] به منظور خوشه‌بندی گراف‌های ویژگی‌دار معرفی می‌کند. روش آن‌ها ابتدا اطلاعات ساختاری که توسط ماتریس

[۲] 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 خوشه بندی در هر دو فضا اقدام به نزدیک کردن این دو ماتریس به کمک تابع هدف می‌کند که به صورت مقابل تعریف شده است:

$$J_{of} = min||A - VV^T||_F^2 + \alpha||VV^T - UU^T||_F^2 + ||F - UU^T F||_F^2 \qquad (۱.۳)$$
$$s.t.\ V \geq 0,\ U \geq 0,\ V^T V = I,\ U^T U = I.$$

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 بهینه سازی 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 [۳۸] ، 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

---

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

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

$$\bar{X} = (I - ۱/۲L)^k X \tag{۲.۳}$$

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. $L = I - A$ که L ماتریس لاپلاسی نرمال شده[۱] می‌باشد. در مرحله بعد برای اعمال خوشه‌بندی طیفی، نیاز به محاسبه ماتریس شباهت بین گره‌ها است که به صورت مقابل عمل کرده‌اند.

$$\min_S ||\bar{X}^T - \bar{X}^T S||_F^۲ + ||S - f(A)||_F^۲ \tag{۳.۳}$$

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 هایپر پارامتر K است که به طور مستقیم بر خروجی الگوریتم تاثیر می‌گذارد که توسط پژوهش دیگری که توسط ژانگ و همکارانش [۳۹] 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 خوشه‌بندی از یکدیگر مستقل‌اند در نتیجه نمی‌توان اطمینان داشت که بازنمایی‌های ایجاد شده برای وظیفه موردنظر (در اینجا خوشه‌بندی) مناسب هستند و همچنین نمی‌توان الگوریتم بازنمایی را بر اساس خطای خوشه‌بندی به طور مناسب به روزرسانی نمود. از این روی، وانگ و همکاران [۴۰] یک روش خوشه

[۱] Normalized Laplacian matrix

بندی یکپارچه توجه محور بر پایه شبکه عصبی گراف ارائه داده‌اند که مرحله بازنمایی 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 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 خوشه براساس مقدار اطمینان تعلق به خوشه به‌روزرسانی 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 خوشه بندی تکرار می‌کنند. معروف ترین الگوریتم این دسته الگوریتم

---

۱ Graph attentional
۲ Decoder
۳ Encoder
۴ Network-based methods
۵ Divisive methods

خوشه‌بندی مارکوف[۱] [۴۱] 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 [۴۲] یکی از الگوریتم‌های این دسته است که از امبدینگ 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 [۴۳] است. این الگوریتم ابتدا پروتئین‌های با درجه بالاتر را به عنوان پروتین‌های هسته [۵] (پروتین‌های آغازین) در نظر گرفته 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[۶] [۴۴] 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.

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 [۴۵]، 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 [۴۶] رابطه‌ی بین شکل گیری مجموعه‌های پروتئینی و هم بیانی پروتئین‌ها را نشان داده است.

- روش‌های هسته‌ـاتصال [۱]: روش‌های هسته‌ـاتصال 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 [۴۷] 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 [۴۸] 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. اخیرا نیز الگوریتم CO-DPC از این دسته بندی ارائه شده است که از نمایه بیان ژنی 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 [۴۹] 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

---

۱ Core-attachment

۲ Functional-information-based

۳ Maximal clique

امبدینگ [۵۰] [۵۱] و تجزیه ماتریسی [۵۲] 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 پروتئین‌ـ 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] [۵۳]، طیف‌سنجی جرمی[۴] [۵۴] به‌شدت غنی شده‌اند. همچنین، روش‌های متن کاوی[۵] برای ایجاد شبکه‌های PPI نیز به صورت گسترده استفاده شده‌اند [۸] [۵۵] [۵۶]. 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 [۵۷]، DIP [۵۸]، Collins [۵۹] و MIPS [۶۰] 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 پروتئین‌ـ پروتئین با در نظر گرفتن ویژگی‌های زیستی و بیان ژنی ایجاد می‌کند که هیمنطور 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 ویژگی‌دار 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:

$$F = GNN_\theta(A, X) \tag{۱.۴}$$

$$\tilde{A} = A + I_N \tag{۲.۴}$$

$$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij} \tag{۳.۴}$$

$$\hat{A} = \tilde{D}^{-۱/۲}\tilde{A}\tilde{D}^{-۱/۲} \tag{۴.۴}$$

$$F = ReLU(\hat{A}ReLU(\hat{A}XW^{(۱)})W^{(۲)}) \tag{۵.۴}$$

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

- استفاده از لایه نرمال‌سازی دسته‌ای بعد از لایه اول گراف کانولوشن
- اعمال regularization $L_۲$ بر روی ماتریس وزن‌ها ($W^{(۱)}$ و $W^{(۲)}$)

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:

$$A_{uv} \sim Bernoulli(۱ - e^{-F_u F_v^T}) \tag{۶.۴}$$

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

$$P(A|F) = \prod_{A_{uv} \in E} ۱ - e^{-F_u F_v^T} \times \prod_{A_{uv} \notin E} e^{-F_u F_v^T} \tag{۷.۴}$$

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

$$-logp(A|F) = -\sum_{A_{uv} \in E} log(۱ - exp(-F_u F_v^T)) + \sum_{A_{uv} \notin E} F_u F_v^T \tag{۸.۴}$$

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

$$L(F) = -\sum_{A_{uv}\in E} log(۱ - exp(-F_u F_v^T)) + \sum_{A_{uv}\in E} F_u F_v^T \qquad (۹.۴)$$

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.

$$L(F) = -E_{(U,V)\sim P_E}[log(۱ - exp(-F_u F_v^T))] + E_{(u,v)\sim P_N}[F_u F_v^T] \qquad (۱۰.۴)$$

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:

$$\theta^* = argmin_\theta \;\; L(GNN(A, X)) + \lambda_۱ ||W^{(۱)}||_۲ + \lambda_۲ ||W^{(۲)}||_۲ \qquad (۱۱.۴)$$

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 می گیریم:

$$F_{uc} = \begin{cases} ۱ & \text{if } F_{uc} > \varphi \\ ۰ & \text{otherwise} \end{cases} \qquad (۱۲.۴)$$

کتاب‌نامه

[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, 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.

[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. 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.

[۳۰] Esmail. Babalian، *Mathematics Discrete in Topics*. Mobtakeran، Tehran، ۲۰۰۷.

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

[32] 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.

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

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

[35] Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] Patra, Sabyasachi and Mohapatra, Anjali. Protein complex prediction in interaction network based on network motif. *Computational Biology and Chemistry*, 89:107399, 2020.

[43] Nepusz, Tamás, Yu, Haiyuan, and Paccanaro, Alberto. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*, 9(5):471–472, 2012.

[44] 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.

[45] Hu, Allen L and Chan, Keith CC. Utilizing both topological and attribute information for protein complex identification in ppi networks. *IEEE/ACM transac-*

tions on computational biology and bioinformatics, 10(3):780–792, 2013.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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

مثال ..... Example

مدول ..... module

واژه‌نامه فارسی به انگلیسی

مثال ..... Example

مدول ..... module

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 18, 2025