دانشگاه صنعتی اصفهان

دانشکده مهندسی برق و کامپیوتر

پیدا کردنِ موتیف در شبکه‌های وزن‌دار برهم‌کنش پروتئین-پروتئین

پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر- هوش مصنوعی

مهسا ایمانی

استاد راهنما

دکتر نیلوفر قیصری

دکتر مهدی صادقی

۱۳۹۰

به نام خدا

دانشگاه صنعتی اصفهان

دانشکده برق و کامپیوتر

پیدا کردنِ موتیف در شبکه‌های وزن‌دار برهم‌کنش پروتئین-پروتئین

پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر- هوش مصنوعی

مهسا ایمانی

استاد راهنما

دکتر نیلوفرقیصری

دکتر مهدی صادقی

۱۳۹۰

نمی دانم چطور با کلمات سپاس خدایی را بگویم که دست مهربان یاری‌دهنده‌ش را در تمام مراحل زندگی‌ام احساس کردم و حتی قبل از آن‌که قدمی برای انجام کاری بردارم او صد قدم برای کمک به من برمی‌داشت. خدا را سپاس می‌گویم سپاسی که بر سپاس های دیگر برتری داشته باشد مانند برتری که پروردگار نسبت به آفریدگان دارد.

از اساتید راهنمایم، سرکار خانم دکتر قیصری و جناب آقای دکتر صادقی که علاوه بر اینکه استاد راهنما و راهگشای من بودند از نظر اخلاقی هم استاد و الگوی من بودند کمال تشکر را دارم. از جناب آقای دکتر موسوی که افتخار شاگردی ایشان را داشتم و ایشان همیشه چه از نظر شیوه تدریس و چه شیوه برخورد با دانشجو و اخلاقی استاد و الگوی من بودند و سر کار خانم دکتر عمومی که داوری این پایان نامه را بر عهده گرفتند، به واسطه پیشنهادهای بسیار پربارشان برای بهبود کار تشکر می‌نمایم.

بر خود لازم می‌دانم از مادرم که اگر موفقیتی در زندگی‌ام کسب کردم آن را مدیون فداکاری‌ها و دلسوزی هایشان هستم، خواهر و برادرانم برای همدلی‌ها و همراهی‌هایشان و پدرم که استقلال فکری و عملی را به من آموخت، خالصانه تشکر و قدردانی کنم.

از سرکار خانم نکویی و سرکار خانم مظاهری، پرسنل مهربان و خوش خلق دفتر تحصیلات تکمیلی دانشکده وجناب آقای دکتر سید محمود مدرس هاشمی رئیس تحصیلات تکمیلی دانشکده برق و کامپیوتر نیز بابت زحماتشان در طول دوره تشکر و قدردانی می‌کنم.

از تمامی دوستان عزیزم در آزمایشگاه هوش مصنوعی و همکلاسی‌های خوبم که حضورشان همواره دلگرم کننده بود، صمیمانه تشکر می‌کنم. بر خود لازم می‌بینم که از عصمت پاکیزه عزیز که همکاری‌اش در کارهای علمی مشترکمان باعث افتخار و نشاط بود و نیز برای یک دنیا مرام و دوستی، از مژگان قربانی و هدی نوری خواهران و مشاوران عزیزم که همیشه از من حمایت می‌کردند، از اعظم عموزادی نازنین، دوست و خواهری که همیشه درک می‌کرد ، از پروین رزاقی عزیز که لحظات زندگی‌ام در خوابگاه را شیرین کرد و از جیمبول برای لبخند زیبای الهام‌بخشش، خالصانه تشکر کنم.

مهسا ایمانی

اردیبهشت ماه ۱۳۹۰

کلیه‌ی حقوق مادی مترتب بر نتایج مطالعات، ابتکارات و نوآوری‌های ناشی از تحقیق موضوع این پایان‌نامه (رساله) متعلق به دانشگاه صنعتی اصفهان است.

تقدیم به بال و پر زندگیم،

مادر نازنینم

که وجودش سمبل عشق و دوست داشتن است،

به پاس همراهی های خالصانه، فداکاری ها، دلواپسی ها و آرزوهای بلندشان برای من.

فهرست مطالب

عنوان **صفحه**

چکیده

شبکه‌های زیست‌شناسی، که عموماً شبکه‌هایِی پیچیده و وسیع هستند، حاوی اطلاعات مهمی می باشند. امروزه نظر به پیشرفتِ قابل ملاحظه آنها ، عملیاتی بر روی این گونه شبکه ها برای استخراجِ اطلاعاتی که هر کدام از آنها دربردارند اِعمال می‌شود. از میان مهم-ترین آنها پیدا کردن موتیف های شبکه است. موتیف به صورت الگوهایی از ارتباطات تعریف می شود که در شبکه با فراوانی بیشتری نسبت به حالت تصادفی مشاهده می شود. تحقیقات انجام شده نشان می‌دهند که چنین ساختارهایی می‌توانند از لحاظ فعالیت در شبکه دارای اهمیت باشند. تاکنون، الگوریتم های مختلفی برای پیدا کردن زیرگراف های با فراوانی بالا در شبکه‌های بدون وزن ارائه شده که دو محدودیت اساسی دارند. اول آنکه تقریبا تمام این روش ها فقط قادر به پیدا کردن موتیف در شبکه‌های بدون وزن هستند. با توجه به رشد شبکه ها در طول دهه اخیر، بررسی خصوصیات دیگری نظیر ناهمگونی یال‌ها که ورای خصوصیات توپولوژیکی آنها است، نیز حائز اهمیت می‌باشد. بررسی شبکه های وزن دار به جای شبکه های بی وزن اجازه می دهد ناهمگونی یال ها در شبکه ها نیز مورد بررسی قرار گیرد. یکی از دلایل اصلی فقدان چنین مطالعاتی آن است که در نظرگرفتن وزن یال‌ها چالش محاسباتی جدیدی را ایجاد می‌کند. محدودیت دیگر روش های موجود آن است که در شمارش زیرگراف ها فقط زیرگراف‌های القایی در نظر گرفته شده است. در نظر گرفتن زیرگراف های غیر القایی یک موتیف شبکه در شبکه‌های برهم‌کنش پروتئین، مسئله‌ای چالش برانگیز وکاملا مطلوب است زیرا این شبکه ها تا کامل شدن و عاری از خطا بودن فاصله بسیار دارند. برهم‌کنش های گزارش شده توسط این شبکه ها شامل برهمکنش هایی می شوند که به طور اشتباه تشخیص داده شدند و نیز بسیاری از برهمکنش‌ها تشخیص داده نشده اند. بنابراین وقوع یک موتیف شبکه خاص دریک شبکه ممکن است شامل یال های اضافه در وقوعش در شبکه دیگری باشد و برعکس. یکی از دلایلی که تحلیل الگوهای گراف‌ها را دشوارتر می کند این است که آنها ذاتا به شکل بردار نیستند و به سادگی نیز قابل تبدیل به بردار نیستند. در این پایان نامه روش جدیدی برای حل دو مشکل مطرح شده در روش‌های قبل پیشنهاد داده شده است که بر خلاف روش های قبلی ، که هنگام نسبت دادن زیرگراف های یافت شده به گروه های مشابه به بررسی یک ریختی زیرگراف ها (انطباق دقیق) می پردازند، از تطبیق غیر دقیق گراف استفاده می کند. تطبیق غیر دقیق زیرگراف ها در شبکه‌های برهمکنش موجود که دارای میزان زیادی خطا هستند کاملا مطلوب است. روش پیشنهادی موتیف هایی وزن دار را معرفی می کند که وزن هر یال نشان دهنده امید حضور آن یال در موتیف اجماع حاصل است. برای این منظور با استفاده از چندجمله ای های متقارن ویژگی های طیفی زیرگراف‌ها بدست آورده می شود و سپس به انطباق وتحلیل بردارهای ویژگی حاصل با استفاده از خوشه بندی به جای طبقه بندی زیرگراف‌ها می پردازد. نتایج تجربی نشان می دهد دخیل کردن وزن ها می تواند سبب پیدا کردن موتیف و ضدموتیف هایی متفاوت از آنچه توسط روش های موجود بدست می-آید، شود. این‌طور به نظر می‌رسد که ذات ارتباط بین وزن‌ها کلی نیست و وابسته به عملکرد شبکه است.

کلمات کلیدی: ، شبکه های وزن دار برهم‌کنش پروتئین، موتیف وزن دار، انطباق غیر دقیق گراف، زیرگراف غیر القایی، ویژگی های طیفی گراف، خوشه بندی

فصل اول
فصل اول: مقدمه

۱ ۱ معرفی

پیدا کردنِ موتیف های شبکه، یک از مسائل مطرح در علم شبکه است و تاکنون روش‌های مختلفی برای آن ارائه شده است. در این پایان‌نامه قصد داریم به پیدا کردنِ موتیف های شبکه های وزن دار برهمکنش پروتئین-پروتئین بپردازیم. این شبکه‌ها یکی از موضوعات اصلی در پروتئومیکس و زیست‌سیستم هستند. بدلیل روش‌های تولید این برهمکنش‌ها این شبکه‌ها دارای خطای بالا هستند و تا کامل شدن فاصله بسیار دارند. این امر پیداکردن موتیف‌های شبکه را بسیارپیچیده‌تر می‌کند.

۱ ۲ پروتئومیکس

توجه به جزئیات به تنهایی نمی تواند پاسخگوی سئوالات بشر و جهان پیچیده باشد همچنان که باید جزئیات را دید باید نگاهی نیز به تصویر کلی حیات داشت در چند سال گذشته دانشمندان متوجه شده اند برای مشاهده تصویری بزرگ از هستی باید علاوه بر کشف ذرات دیدی کلی به جهان داشت. به دانش نگرش سیستمیک به حیات و پدیده های مرتبط زیستی، زیست‌سیستم گفته می‌شود.

یکی از کاربردهای اصلی این رشته که امروزه هزینه ها و تمرکز زیادی را به خود اختصاص داده است، استفاده از سیستم بیولوژی در مدل سازی اثر دارو ها در سلول های هدف و مسیرهای بیوشیمیایی درگیر است.

دانشمندان امیدوارند که در آینده نزدیک با مدل سازی های دقیق‌تر بتوانند دارو های اختصاصی تر با اثرات جانبی کمتر و بهینه را طراحی کنند.

موضوعات اصلی در زیست سیستم شامل آنالیز شبکه های مختلف است که مهمترین این شبکه ها عبارتند از[۱و۲]:

- شبکه های ژنی
- شبکه‌های برهمکنش پروتئین-پروتئین (ب.پ.پ)[1]
- شبکه های متابولیسمی
- شبکه های سیگنالی

برای مدل سازی هر کدام از این شبکه ها با توجه به ویژگی هایی که هر کدام از آنها دارند از روش ها و الگوریتم های خاصی استفاده می شود ولی آن چه مهم است این است که در نهایت هدف زیست‌سیستم رسیدن به یک مدل واحد و کشف ارتباطات بین این شبکه ها و نقش آنها در حیات سلول ، ارتباط بین سلول ها و شرایط محیطی و در نهایت پاسخ موجود به تغییرات محیطی است.

ژن‌ها با عملکرد پروتئین‌هایی که می‌سازند نقش خود را اعمال می‌کنند. بنابراین لازمه شناخت کامل ژن‌ها ، شناخت کامل پروتئین‌ها است. در واقع هم‌اکنون مطالعه پروتئین‌ها با توجه به اهمیت آنها و پیشرفت‌های سریع ابزاری و تکنیکی در این رشته، جایگاه خاصی یافته است. با تکمیل طرح ژنوم انسان مشخص شد که مکانیسم ملکولی رفتار سلول‌ها در شرایط مختلف را نمی‌توان از روی توالی ژن‌های آنها پیشگویی کرد. رفتار سلولی و تمام فعالیت‌هایی که در سلول انجام می‌شود به عهده پروتئین‌های بیان شده به وسیله ژن است. در واقع برای ارتباط ژنوم با رفتار سلول‌ها باید پروتئین‌های سلول‌ها را شناخت [۱و۲].

بنابراین برای شناسایی مکانیسمهای مولکولی رفتار سلولی و واکنشهای زیستی، لازم است پروتئینهایی که در یک سلول بیان می‌شود، تغییرات آنها در شرایط مختلف، عملکرد آنها و همچنین برهمکنش‌های بین پروتئینهای مختلف در یک سلول، بررسی شود. به مجموعه این بررسیها، نقشه برداری پروتئوم یا پروتئومیک ، گفته میشود. امروزه پروتئومیک و ژنومیک دو فعالیت برجسته‌ای هستندکه بیوانفورماتیک دانان به آن مشغول هستند [۱و۲].

ژنومیک شامل تجزیه و تحلیل داده‌ها و اطلاعات ژنتیکی ژنوم موجودات است. پروتئومیک به آنالیز پروتئینهای یک موجود زنده گفته می‌شود. سنجش مقایسه ای پروتئین ها به صورت جامع در یک مقیاس وسیع موضوع علم پروتئوم یا پروتئومیکس می باشد. از دیدگاه بیوشیمیست ها پروتئومیکس، مطالعه بیش از یک پروتئین

[1] Protein-protein interaction (PPI)

در یک زمان خاص است. اولین مطالعات بر روی پروتئین در سال ۱۳۵۴ با ابداع ژل دو بعدی آغاز شد. علم پروتئومیکس بسیار گسترده است. شناخت پروتئین ها، بررسی کمی آنها در سلول ها، بافت ها و مایعات زیست شناختی، سنجش تغییرات در بیان پروتئین ها در سلول های بیمار در مقابل سلول های طبیعی، توصیف تغییرات پس از ترجمه، مطالعه برهمکنش های پروتئین-پروتئین، تعیین موقعیت ، شناسایی عملکرد سلولی در سطح پروتئین ها، شناسایی ژن‌های ناشناخته به کمک پروتئین‌ها، بسیاری از کاربردها و جوانب دیگر آن می باشند. از مهمترین اهداف تحقیقات پروتئومیکس نیز شرح و توصیف مکانیسم های مولکولی دخیل در فرایندهای سلولی، ویژگی شبکه های پیچیده پروتئینی و اختلال در آنها، کشف بیومارکرهای پروتئینی برای آشکار سازی و تشخیص بیماری-ها و شناخت اهدافی برای طراحی درمآنهای داروئی می باشند. در مقابل ژنوم که نسبتا ثابت و بی تغییر است، پروتئومیکس دینامیک و متغیر است. بنابر این، می توان پروتئومیکس را علم "پس از ژنوم" نامید، چرا که پروتئومیکس نقش قابل توجهی در عبور از ژنومیکس به کاربردهای سودمند بالینی به ویژه در عرصه های تشخیص و پیشگیری دارد [۱و۲].

به دلیل اهمیت پروتئومیک، دهه آینده در بین بیولوژیستهای مولکولی به دهه پروتئومیک مشهور شده است. حجم مقالاتی که در چند سال اخیر در مورد پروتئومیک انتشار یافته است در این سال‌ها با شتاب بسیار زیاد، در حال افزایش است.

۱ ۳ علم شبکه

علم شبکه، مطالعه‌ی سیستم‌های پیچیده (تشکیل شده از بخش‌های در حال تعامل) از دید تئوری گراف است. ریشه‌ی علم شبکه به مطالعه‌ی شبکه‌های اجتماعی در دهه‌ی ۱۳۳۰ باز می‌گردد. علم شبکه بینش وسیعی در مورد چگونگی عملکرد اجزای مختلف جهان ما، از سلول تا اینترنت و اقتصاد فراهم کرده است. در دهه اخیر به-دلیل رشد روز افزون میزان دسترسی به داده‌های شبکه نظیر شبکه جهانی اینترنت[2]، شبکه‌های برهمکنش پروتئین، شبکه‌های غذا و نیز میزان دسترسی به کامپیوترهای شخصی قدرتمند برای پردازش داده، علم شبکه مورد توجه زیادی قرار گرفته است. امروزه علم شبکه تقریبا مورد توجه محققان تمامی رشته‌های علمی و صنعتی است.

رفتار سلولی و تمام فعالیت‌هایی که در سلول انجام می‌شود بر عهده پروتئین‌ها است. همه پروتئین‌ها با هم برهم‌کنش دارند و تقریباً می‌توان گفت که همه پروتئین‌ها اثر خود را با همکاری پروتئین‌های دیگر در سلول اعمال می‌کنند و هیچ پروتئینی نیست که در یاخته به تنهایی عمل کند. اینکه کدام پروتئین در کدام زمان، و در کدام

[2] WWW

سلول ایجاد شود، رویکردی است که براساس مکانیسم های تنظیم سلولی و مبتنی بر مقتضیات و نیازهای (کوتاه مدت، میان مدت و درازمدت) سلول شکل می گیرد. این رخدادها، به نوبه خود، نیز ویژگی های متفاوت سلول های یک موجود منفرد را به روشنی تبیین می کند. لذا برای بررسی مکانیسم عمل پروتئین ها در سلول از شبکه های برهمکنش پروتئین-پروتئین که نحوه ی تعامل پروتئین ها را نشان می دهند، استفاده می شود.

۱ ۳ ۱ شبکه‌های برهمکنش پروتئین-پروتئین

فرآیندهای سلولی به‌وسیله تعامل پروتئین‌ها با یکدیگر، با محیط و با DNA تعریف می‌شوند. فهمیدن این-که چطور پروتئین‌ها با یکدیگر در سطح شیمیایی تعامل می‌کنند، گام اولیه برای مدل‌کردن فرآیندهای سلولی است. فهم این موضوع می‌تواند مفاهیم مهمی برای درک سیستم‌های بزرگتر آشکار شده به‌وسیله شبکه‌های برهمکنش پروتئین-پروتئین (ب.پ.پ) و شبکه‌های تنظیمی ژن[3] را بیان کند. در شبکه‌های برهمکنش پروتئین، گره‌ها پروتئین‌ها و یال‌های غیرجهت‌دار، برهم‌کنش‌های خروجی پروتئین-پروتئین را نشان می‌دهند. در شبکه‌های تنظیمی ژن، گره‌ها ژن‌ها و پروتئین‌ها ، و یال‌های جهت‌دار، تولید یک پروتئین به‌وسیله‌ی یک ژن یا تنظیم یک ژن به‌وسیله‌ی یک پروتئین را نشان می‌دهد.

در چند سال گذشته شبکه‌های ب.پ.پ ، کاربردهایی نظیر پیش‌بینی برهم‌کنش‌های دامنه-دامنه[4] و پیش-بینی برهم‌کنش‌های پروتئین- پپتید[5] را براساس موتیف‌های شبکه یافته‌اند. از وابسته‌سازی یک شبکه ب.پ.پ با یک شبکه پروتئین- DNA می‌توان برای تفسیر ژن‌هایی که قبلا طبقه‌بندی نشده‌اند، استفاده کرد.

۱ ۳ ۲ موتیف‌های شبکه

پیمانه ای بودن[6] روشی استاندارد در طراحی سیستم‌ها و مهندسی می‌باشد. ساختار پیمانه ای امکان استفاده‌ی مجدد از زیربخش‌های متداول را فراهم می‌کند. مهندسان به منظور کنترل و مدیریت پیچیدگی سیستم‌های بزرگتر از سازمان‌دهی سلسله مراتبی استفاده می‌کنند. علاوه بر این در علم شبکه، این ویژگی در شبکه‌هایی که به طور طبیعی در حال رشد و تکامل هستند، نیز شایع است. مطالعه‌ی چنین زیرشبکه‌هایی که به طور طبیعی رخ داده‌اند، منجر به ایجاد بینشی در مورد نقش مجموعه‌ای از گره‌های یک شبکه در پردازش اطلاعات شده است [۳].

---

[3] Gene Regularly Networks
[4] Domain-Domain interactions
[5] De Novo Protein-Peptide Interactions
[6] Modularity

موتیف‌های شبکه برای درک پیمانه ای بودن ساختار کلی شبکه‌ها مفید هستند. موتیف‌ها برای اولین بار توسط میلو و همکارانش معرفی شدند [۳] و اهمیت آنها به عنوان واحد پردازش اطلاعات از نظر تئوری مدل شده است [۴و۵]. هم‌چنین موتیف‌ها به صورت تجربی مورد آزمایش و بررسی قرار گرفته‌اند [۶-۹].

موتیف‌های شبکه به‌صورت زیر شبکه‌هایی که مهم و غیرتصادفی هستند، تعریف می‌شوند [۳]. موتیف شبکه به طور معمول از تکرار زیاد آن در شبکه واقعی در مقایسه با نسخه‌های تصادفی همان شبکه، تعیین می‌شوند [۳و۴]. به همین ترتیب، ضد موتیف از تکرار کم آن در شبکه اصلی در مقایسه با نسخه‌های تصادفی همان شبکه تعیین می‌شوند.

مطالعات اخیر در زمینه هم‌تراز کردن[7] شبکه‌های زیست‌شناسی، زیر شبکه‌های حفظ شده را آشکار کرده است [۱۰].

پیدا کردن همه‌ی نمونه‌های یک گراف معین به عنوان زیرگراف یک شبکه از جهات زیر لازم است:

- تعیین این نکته که آیا یک گراف معین (شاید به صورت تجربی تعیین شده باشد) یک موتیف معنادار هست.
- بررسی خوشه‌بندی موتیف به منظور تعیین این که چطور موتیف‌ها می‌توانند بخش‌هایی از ساختارهایی بزرگتر باشند و تعیین این نکته که معناداری موتیف چقدر به دقت شبکه وابسته است.
- پیداکردن موتیف‌های مبتنی بر فراوانی با پیداکردن موتیف‌های مبتنی بر بقا به روش‌های خاص (هم‌تراز کردن شبکه) ترکیب شود.

۱ ۴ تبیین مساله پایان نامه

امروزه شبکه های بزرگ، مانند شبکه های زیست‌شناسی که هر کدام شامل هزاران گره می باشند، بسیار مورد توجه محققان علم پروتئومیکس قرار گرفته اند. یکی از تلاش‌هایی که در این زمینه مورد پی گیری قرار گرفته است، پیدا کردن موتیف در این شبکه ها از جمله شبکه های برهم‌کنش پروتئین-پروتئین است. موتیف، به صورت الگوهایی از زیرگراف ها تعریف می شود که در شبکه با فراوانی بیشتری نسبت به حالت تصادفی مشاهده می شود. به نظر می رسد این چنین ساختارهایی از لحاظ فعالیت در شبکه دارای اهمیت باشند. الگوریتم‌های کنونی برای پیدا کردن موتیف های با اندازه ی بزرگ با مشکل مواجه شده اند. از طرفی داده های موجود برای شبکه های برهم-کنش پروتئین-پروتئین خالی از خطا نیستند و هیچ یک از روش‌های پیش‌بینی و استخراج برهم‌کنش‌ها در شبکه، به

[7] Graph Alignment

شبکه واقعی همگرا نمی‌شود. شبکه‌های گزارش شده توسط پایگاه داده های مختلف برای یک ارگانیسم دارای هم پوشانی بسیار کمی هستند. بنابراین این شبکه ها خالی از خطا نیستند و نیاز به منابع اطلاعاتی دیگر برای یکپارچه سازی داده[۸] و وزن‌دار کردن شبکه‌های برهمکنش پروتئین وجود دارد. به‌همین دلیل امروزه علاقه به استفاده از شبکه‌های برهم‌کنش وزن‌دار پروتئین به جای شبکه‌های بی‌وزن رو به افرایش است. در این شبکه‌ها وزن هر یال نشان‌دهنده میزان اعتبار آن است.

تاکنون، الگوریتم های مختلفی برای پیدا کردن زیر گراف های با فراوانی بالا در مجموعه شبکه ها یا در یک گراف ارائه شده است. این الگوریتم ها یا تنها به شمارش انواع خاصی از زیرگراف ها می پردازند و یا شرطی بر روی اندازه ی زیرگراف دارند. در تکنیک‌های فعلی بدلیل محدودیت اندازه، موتیف‌های با حداکثر ۸ یا ۹ گره را پیدا می‌کنند. می توان این تکنیک‌ها را دو دسته تقسیم کرد، دسته اول به شمارس کامل زیرگراف‌ها پرداخته و دسته دوم از نمونه‌گیری زیرگراف استفاده می‌کند.

الگوریتم‌های موجود دو محدودیت اساسی دارند. اول آنکه تقریبا تمام این روش ها فقط قادر به پیدا کردن موتیف در شبکه های بدون وزن هستند. با توجه به رشد شبکه ها در طول دهه گذشته، بررسی خصوصیات دیگری نظیر ناهمگونی یال‌ها که ورای خصوصیات توپولوژیکی آنها است، نیز حائز اهمیت می‌باشد. بررسی شبکه های وزن دار به جای شبکه های بی وزن اجازه می دهد ناهمگونی یال ها در شبکه ها نیز مورد بررسی قرار گیرد. یکی از دلایل اصلی فقدان چنین مطالعاتی بر روی شبکه های وزن‌دار آن است که در نظر گرفتن وزن یال‌ها چالش محاسباتی جدیدی را ایجاد می‌کند.

محدودیت دیگر اکثر روش های موجود آن است که در شمارش زیرگراف‌ها فقط زیرگراف‌های القایی در نظر گرفته شده است. در روش‌های پیشنهاد شده جدیدتر برای شبکه‌هایی با خطای بالا درنظر گرفتن زیرگراف های غیر القایی علاوه بر زیرگراف‌های القایی پیشنهاد شده است. در نظر گرفتن زیرگراف های غیر القایی یک موتیف شبکه در شبکه‌های برهمکنش پروتئین، مسئله‌ای چالش برانگیز وکاملا مطلوب است زیرا این شبکه ها تا کامل شدن و عاری از خطا بودن فاصله بسیار دارند. برهمکنش‌های گزارش شده توسط این شبکه ها شامل برهمکنش‌هایی می شوند که به طور اشتباه تشخیص داده شدند و نیز بسیاری از برهمکنش ها تشخیص داده نشده اند. بنابراین وقوع یک موتیف شبکه خاص دریک شبکه ممکن است شامل یال های اضافه در وقوعش در شبکه دیگری باشد و برعکس. همچنین یکی از دلایلی که تحلیل الگوهای گراف ها را دشوارتر می کند این است که آنها ذاتا به شکل بردار نیستند و به سادگی نیز قابل تبدیل به بردار نیستند.

[8] Data integration

۱ ۵ اهداف و انگیزه پایان نامه

هدف اصلی این تحقیق واردکردن وزن یال ها در مطالعه ی موتیف های شبکه های برهمکنش پروتئین است زیرا این شبکه ها هنوز با شبکه ای عاری از خطا فاصله ی بسیار دارند. در این شبکه ها وزن هر یال نشان‌دهنده میزان اعتبار آن است. بدین ترتیب ناهمگونی یال ها در شبکه ها نیز مورد بررسی قرار گیرد.

از آنجا که شبکه ها در طول دهه گذشته رشد کرده اند، باید به بررسی خصوصیات دیگری ورای خصوصیات توپولوژیکی آن‌ها پرداخت. یکی از این مطالعات بررسی شبکه های وزن دار به جای شبکه های بی-وزن می باشد که اجازه می دهد ناهمگونی یال ها در شبکه ها نیز مورد بررسی قرار گیرد.

اگر یک ارتباط کلی بین وزن ها و توپولوژی باشد، در نظر گرفتن وزن ها در بررسی سیستم های شبکه های اطلاعات جدیدی فراهم نمی کند. در دسترس بودن شبکه های وزن دار ما را قادر می سازد به بررسی ارتباط بین توپولوژی برهمکنش ها با قدرت این برهمکنش ها به پردازیم. این‌طور به نظر می رسد که ذات این ارتباط کلی نیست و مرتبط با عملکرد شبکه است. در نتیجه دخیل کردن وزن ها در بررسی موتیف ها می تواند درک ما را از سیستم های شبکه ای بهبود دهد.

از سوی دیگر فقط تعداد محدودی مطالعه در زمینه شبکه های زیستی وجود دارد که زیرگراف های غیر القایی را در پیداکردن موتیف ها درنظرگفته است. دلیل آن این است که تعداد بسیار بیشتری زیرگراف غیر القایی یک ریخت با یک توپولوژی خاص وجود دارد و بنابراین شمارش زیرگراف های غیر القایی شبکه بسیار دشوارتر است. در نظر گرفتن زیرگراف های غیرالقایی یک موتیف شبکه نه تنها چالش برانگیز است که کاملا براساس شبکه های برهمکنش فعلی تا کامل شدن و عاری از خطا بودن فاصله بسیار دارند کاملا مطلوب است.

۱ ۶ روش پیشنهادی

یکی از موانع برسر پیداکردن موتیف های شبکه، مسئله ایزومورفیسم گراف است که از نظرمحاسباتی پیچیده و در طبقه مسائل NP-complete قرار می‌گیرد. مسئله تطبیق گراف به دو دسته اصلی تطبیق دقیق گراف[۹] و تطبیق غیردقیق گراف[۱۰] تقسیم می شود. مسئله تطبیق دو گراف به صورت دقیق همان مسئله ایزومورفیسم می باشد که برای بدست آوردن زیرگراف های القایی موتیف‌ها از آن استفاده می‌شود. به عبارت دیگر در انطباق دو زیرگراف تعداد گره ها و یال ها و همچنین طریقه اتصال گره ها به یکدیگر باید یکسان باشد. استفاده از انطباق

[9] exact graph matching
[10] inexact graph matching

دقیق برای این شبکه‌های وزن‌دار بدلیل ناهمگونی بالای یال‌ها امکان‌پذیر نیست. در تطبیق غیر دقیق دو زیرگراف، تعداد گره ها و یال ها و همچنین طریقه اتصال گره ها به یکدیگر می تواند متفاوت باشد و هدف یافتن تطبیقی است که این اختلاف درآن حداقل باشد. تطبیق غیر دقیق گراف در طبقه مسائل NP-complete قرار می گیرد. ازین رو الگوریتمهای زیادی برای ارائه تقریبی از راه حل مطلوب پیشنهاد شده اند [۱۱].

برای رفع دو محدودیت موجود در شمارش موتیف‌های غیر القایی در شبکه های وزن‌دار در این پایان نامه قصد داریم برخلاف روش های قبلی ، به جای استفاده از تطبیق دقیق (ایزومورفیسم) از تطبیق غیر دقیق گراف استفاده کنیم. به عبارت دیگر در انطباق دو زیرگراف تعداد گره ها و یال ها و همچنین طریقه اتصال گره ها به یکدیگر می تواند متفاوت باشد. برای این منظور با تبدیل زیرگراف ها به بردارهای ویژگی با استفاده از ویژگی‌های های طیفی گراف به انطباق وتحلیل آنها می‌پردازیم. تطبیق غیر دقیق زیرگراف ها براساس شبکه های برهمکنش موجود که دارای میزان زیادی خطا هستند کاملا مطلوب است. در الگوریتم پیشنهادی مفهوم موتیف بسط داده می شود تا وزن یال ها که نشان دهنده ی قدرت برهمکنش(در مورد شبکه های برهمکنش پروتئین ها نشان دهنده ی قابلیت اعتماد برهمکنش ها) میان دو عنصر سازنده است در نظرگرفته شود.

یکی از دلایلی که تحلیل الگوهای گراف ها را دشوارتر می کند، آن است که آنها نه در ذات به شکل بردار هستند و نه به سادگی به بردار تبدیل می شوند. ازاین رو الگوریتم پیشنهادی برای تطبیق گراف ها،یک بردار از ویژگی های طیفی آنها می سازد. ویژگی های طیفی که ساخته می شود با تغییر جایگاه گره ها در ماتریس مجاورت تغییر نمی کنند و از تمام ماتریس طیفی بهره می‌برد. روش مورد استفاده از مقادیر چندجمله ای های متقارن[11] برای ساخت بردار ویژگی گراف استفاده می‌کند. سپس به خوشه‌بندی بردارهای ویژگی بدست آمده با استفاده از روش خوشه بندی جدید K-means++ می‌پردازد. سپس برای تعیین معناداری موتیف های اجماع بدست آمده به جای استفاده از تعریف Z-score ساده که در الگوریتم‌های پیشین استفاده می‌شد از تعمیم Z-score که براساس وزن یال‌های زیرگراف است استفاده می‌شود.

۱ ۴ دستاوردها

روش پیشنهادی، اولین روش جامع در زمینه مطالعه‌ی موتیف‌های شبکه‌های وزن‌دار است که محدودیتی برروی نوع زیرگراف‌های مورد بررسی ندارد و همزمان قادر به حل هر دو مشکل موجود در زمینه پیدا کردن

[11] symmetric polynomials

موتیف‌ها، یعنی دخیل کردن وزن یالها و درنظر گرفتن زیرگراف های غیرالقایی که در شبکه ای با خطای بالابسیار مورد نیاز است، می‌باشد.

این روش برای اولین بار از مفهوم انطباق غیر دقیق در انطباق زیرگراف ها به جای انطباق دقیق استفاده می-کند. روش پیشنهاد شده برای انطباق غیر دقیق قابلیت اعمال به بقیه مسائل مطرح در زمینه شبکه های وزن‌دار مانند یافتن ترکیبات پروتئینی را نیز دارد.

در روش‌های قبلی، موتیف یافت شده یا وجود دارند یا ندارند. اما روش پیشنهادی مفهوم موتیف را ازحالت وجود داشتن یا نداشتن به طیفی از وجود داشتن بسط می‌دهد. در موتیف ارائه شده وزن هر یال نشان‌دهنده‌ی امید حضور آن یال در موتیف است.

۱ ۸ روند ارائه ی مطالب

ازآنجاکه مطالعه برهمکنش های پروتئین-پروتئین یکی از موضوعات مهم در زیست سیستم و علم پروتئومیکس است ، در فصل دوم مفاهیم و روش های کنونی گرداوری این برهمکنش ها بیان می شود. در فصل سوم، بدلیل آنکه بخش بزرگی از داده های کنونی برهمکنش های پروتئین-پروتئین، دارای خطا می باشند و دراین پایان نامه از شبکه های یکپارچه شده وزن‌دار استفاده می شود، مفهوم یکپارچه سازی بانک های اطلاعاتی و اهمیت آن در حوزه ی داده های زیست شناختی بیان می شود. سپس در فصل چهارم شبکه ها، مفهوم و اهمیت موتیف ها در شبکه ها و کارهای مرتبط در زمینه یافتن موتیف ها در شبکه ها از جمله در شبکه های برهمکنش های پروتئین-پروتئین بیان می شود. در فصل پنجم الگوریتم پیشنهادی و در فصل ششم نتایج تجربی آورده می شود. سپس از مطالب گفته شده در فصل آخر نتیجه گیری و پیشنهاداتی ارائه می‌شود.

فصل دوم
پیش زمینه زیست شناختی

۲ ـ ۱ مقدمه

در این فصل اطلاعات پایه در مورد پروتئین، برهم‌کنش‌های پروتئین-پروتئین و شبکه‌های برهم‌کنش پروتئین ارائه می‌شود. برهم‌کنش پروتئین-پروتئین به ارتباط مولکول‌های پروتئین با یکدیگر برمی‌گردد. این ارتباطات از چندجهت مانند نواحی تحقیقاتی کلی مثل بیوشیمی، بیوفیزیک، فرآیندهای زیست‌شناختی خاص، مسیرهایی از قبیل مسیر انتقال سیگنال[۱۲] و مطالعات در سطح سیستم شبکه‌های ارگانیزم حائز اهمیت است.

۲ ـ ۲ پروتئین و کارکرد آن

پروتئین‌ها مواد آلی بزرگ و یکی از انواع درشت‌ملکول‌های زیستی هستند که از زیرواحدهایی به نام اسید آمینه ساخته شده‌اند. اسیدهای آمینه مثل یک زنجیر خطی توسط پیوند پپتیدی میان گروه‌های کربوکسیل و آمین مجاور به یکدیگر متصل می‌شوند تا یک پلی پپتید را به وجود بیاورند. ترتیب اسیدهای آمینه در یک پروتئین توسط ژن مشخص می‌شود. به جزدر بعضی از ارگانیسم ها ،کد ژنتیک ۲۰ تا اسید آمینه استاندارد را معرفی می‌کند، پروتئین‌ها معمولا به یکدیگر می‌پیوندند تا یک وظیفه‌ای را با یکدیگر انجام دهند که این خود باعث استوار شدن پروتئین می‌شود. چون ترتیب‌های نامحدودی در توالی و طول زنجیره اسید آمینه‌ها در تولید پروتئین‌ها وجود دارد، از این رو انواع بی‌شماری از پروتئین‌ها نیز می‌توانند وجود داشته باشند.

---

[12] Signal Transduction Pathway