# کاربردهای پردازش سیگنال گراف در پیش‌بینی برهمکنش‌های پروتئین-پروتئین

**رکسانا بابائی و آرتا امیر جمشیدی***

**ایران، تهران، دانشگاه تهران، دانشکدگان علوم، دانشکده ریاضی، آمار و علوم کامپیوتر، آزمایشگاه تحقیقاتی پیشرفته سیستم‌های زیستی و سرطان**

تاریخ دریافت: ۱۴۰۲/۰۲/۱۱ تاریخ پذیرش: ۱۴۰۲/۰۶/۲۵

## چکیده

برهمکنش‌های پروتئین-پروتئین، اتصال‌های فیزیکی و شیمیایی بین دو یا چند پروتئین هستند که نقشی حیاتی را در طیف وسیعی از فرایندهای سلولی ایفا می‌کنند. پردازش سیگنال گراف یک ابزار جدید و قدرتمند برای تجزیه و تحلیل گراف‌ها می‌باشد که اخیرا با موفقیت زیادی بر روی شبکه‌های پروتئین-پروتئین اعمال شده است. در این مقاله مفاهیم مقدماتی پردازش سیگنال گراف جهت مدل‌سازی شبکه برهمکنش پروتئین-پروتئین مرور می‌شود. ویژگی‌های گراف با استفاده از تبدیل موجک طیفی گراف و خوشه‌بندی سلسه مراتبی بررسی می‌گردد. در این مقاله تعدادی از الگوریتم‌های مهم پیش‌بینی برهمکنش‌های ناشناس از جمله الگوریتم GRABP، Spectral link و $L_3$ بررسی شده و نتایج حاصل از این الگوریتم‌ها بر روی شبکه‌های گراف برهمکنش پروتئین-پروتئین بدن انسان، گیاه رشادی، کرم الگانس و مخمر گزارش شد. نتایج این الگوریتم‌ها نشان داد که پردازش سیگنال گراف می‌تواند با تقریب مناسبی برهمکنش‌های ناشناس را در شبکه گراف موجودات زنده تشخیص دهد.

**واژه های کلیدی**: شبکه برهمکنش پروتئین-پروتئین، پردازش سیگنال گراف، تبدیل موجک طیفی گراف، یادگیری ماشین.

* نویسنده مسئول، پست الکترونیکی: arta.jamshidi@ut.ac.ir

## مقدمه

گراف یک روش مدل سازی برای نشان دادن انواع مختلف شبکه‌ها مانند شبکه‌های بیولوژیکی، اجتماعی و حسگرها می‌باشد. پردازش سیگنال گراف یک زمینه در حال رشد است که به بررسی روش‌های پردازش سیگنال کلاسیک در داده‌های مدل شده با گراف می‌پردازد. پردازش سیگنال گراف مجموعه‌ای قدرتمند از ابزارها را برای تجزیه و تحلیل داده‌ها در مقیاس گسترده فراهم می‌سازد (۱). برهمکنش‌های پروتئین-پروتئین تعاملات مهمی هستند که برای بسیاری از فرآیندهای بیولوژیکی مانند تنظیم متابولیسم بدن، ضروری می‌باشند. شبکه برهمکنش پروتئین-پروتئین دارای حجم بالایی از داده است که مطالعه این نوع از تعاملات با روش‌های کلاسیک را بسیار دشوار می‌سازد. پردازش سیگنال گراف به دلیل توانایی تشخیص الگوی رفتاری و تقریب میزان ارتباطات شبکه برهمکنش، مزیت بیشتری نسبت به روش‌های پردازش سیگنال کلاسیک دارد. در این مقاله عملکرد پردازش سیگنال گراف و الگوریتم‌های پیش‌بینی برهمکنش پروتئین-پروتئین (۲)، را روی داده‌های مختلف بررسی کرده‌ایم. در بخش نخست این مقاله کاربرد گراف در پیش‌بینی تداخلات دارویی، مساله پیش‌بینی پیوند و کشف دارو را بررسی کرده و سپس اهمیت برهمکنش پروتئین-پروتئین را برای کشف داروها مطالعه کردیم. در بخش پردازش سیگنال گراف، مفاهیم مقدماتی این ابزار بررسی شده و سپس با از استفاده از مفهوم ماتریس لاپلاسین در بخش تبدیل موجک طیفی گراف، ویژگی‌های شبکه برهمکنش پروتئین-پروتئین با توجه به ساختار موضعی داده‌ها استخراج شد (۳). در بخش خوشه‌بندی سلسله مراتبی با الگوریتم پیوند میانگین، با استفاده از ویژگی‌های

بدست آمده از تبدیل موجک، گراف متناظر با شبکه برهمکنش داده‌ها به خوشه‌های مختلف تقسیم شد (۳). در بخش میدان تصادفی مارکوف، یک مدل از این میدان بر روی سیگنال‌های گراف انتخاب شده و با استفاده از این مدل، احتمال وجود یال در گراف متناظر با شبکه برهمکنش داده‌ها بررسی شد (۴). در بخش طرح الگوریتم پیش‌بینی برهمکنش پروتئین-پروتئین، با استفاده از ابزارهای مطالعه شده، یک الگوریتم برای پیش‌بینی برهمکنش‌های ناشناس در گراف تشریح گردید (۵). این الگوریتم شامل یک مساله یادگیری ماشین است، به‌طوری که در این یادگیری با استفاده از توزیع گیبس تعریف شده بر روی میدان تصادفی مارکوف، بیشترین احتمال وجود برهمکنش میان پروتئین‌ها محاسبه می‌شود (۶). در بخش پیاده‌سازی الگوریتم و ارزیابی عملکرد، الگوریتم بررسی شده بر روی مجموعه داده‌های مختلف اجرا شده و با استفاده از معیار منحنی مشخصه عملکرد میزان بازدهی الگوریتم مشخص شد. در بخش آخر، الگوریتم بررسی شده با سایر الگوریتم‌ها مقایسه گردید. نتایج به دست آمده بیانگر دقت قابل قبول الگوریتم پیش‌بینی پیوندها با استفاده از پردازش سیگنال گراف است.

**گراف و پیش‌بینی تداخل داروها:** تداخل دارویی، پدیده‌ای ناخواسته و گاه خطرناک است که در اثر مصرف همزمان دو یا چند دارو رخ می‌دهد. این تداخلات می‌توانند منجر به واکنش‌های جانبی، جراحت و یا حتی مرگ شوند (۷). شناسایی تداخل‌های دارویی بالقوه، نقشی حیاتی در کاهش عوارض جانبی و تسریع روند درمان دارند (۸). قبل از همه‌گیری روش‌های محاسباتی، شناسایی تداخل‌های دارویی بر پایه مطالعه منابع مختلف مانند کتاب‌های مرجع و داده‌های ثبت شده بود. در این روش‌ها، فرض می‌گردید که داروهایی با ویژگی‌های مشابه، تداخل‌های دارویی مشابهی را از خود نشان می‌دهند. در این روش‌ها برخلاف کارایی قابل قبول، نقص‌هایی نیز مشاهده می‌شود. از جمله این نقص‌ها می‌توان به نادیده‌گیری همبستگی بین تداخلات دارویی اشاره کرد. در سال‌های اخیر، نظریه گراف به عنوان ابزاری قدرتمند برای بررسی همبستگی و شباهت داده‌ها، مورد توجه قرار گرفته است (۹). از این نظریه می‌توان برای پیش‌بینی تداخل‌های دارویی استفاده نمود. روش‌های مبتنی بر نظریه گراف به دلیل دقت و کارایی بالا، قابلیت پیش‌بینی تداخل‌های دارویی جدید و بررسی همبستگی بین تداخل‌های دارویی، ارزش بالایی دارند (۱۰). از نظریه گراف می‌توان در یادگیری ماشین استفاده کرد. در این روش، تداخل‌های دارویی با استفاده از الگوریتم‌های مختلف شناسایی، خوشه‌بندی و استخراج می‌گردند (۱۱). از جمله این الگوریتم‌ها می‌توان به الگوریتم‌های جستجو برای یافتن مسیر بین راس‌های گراف و خوشه‌بندی برای دسته‌بندی راس‌های گراف بر اساس میزان شباهت بین راس‌ها اشاره نمود (۱۲).

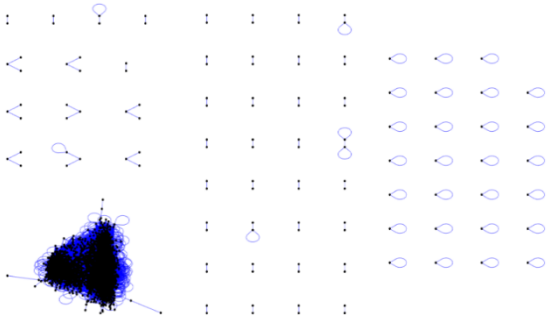
**پیش‌بینی پیوندها و کشف دارو:** بیماری‌ها، پدیده‌های پیچیده‌ای هستند که از روابط غیرخطی بین سلول‌های منفرد تا موجودات زنده را دربرمی‌گیرند (۱۳). برای یافتن پاسخ مناسب به هر بیماری، یعنی طراحی و تولید داروهای مؤثر، باید به مجموعه‌ای از این پدیده‌ها توجه نمود (۱۴). فرایند کشف دارو به طور کلّی روشی پرهزینه و زمان‌بر است که شامل پنج مرحله کشف و تحقیقات پیش‌بالینی، بررسی ایمنی، تحقیقات بالینی، بررسی سازمان‌های نظارتی و نظارت بر ایمنی پس از فروش می‌باشد (۱۵). روش‌های نوین کشف دارو مبتنی بر علم داده هستند. الگوریتم‌های معرفی شده در پردازش داده می‌توانند سرعت فرآیند کشف دارو را به طور قابل توجهی افزایش دهند. در علم داده چهار دسته روش اصلی در زمینه کشف دارو وجود دارد: ۱-رویکردهای مبتنی بر لیگاند، ۲-رویکردهای اتصال، ۳-رویکردهای مبتنی بر شبکه و ۴-رویکردهای مبتنی بر یادگیری ماشین. در روش‌های مبتنی بر لیگاند فرض می‌شود که داروهای مشابه تمایل به اتصال به عوامل بیماری‌زای مشابهی را دارند. از آنجایی که این رویکرد

برای پیش‌بینی نتایج خود از شباهت لیگاندها استفاده می‌کند، این روش برای دستیابی به جواب، نیازمند نمونه‌هایی اولیه از برهمکنش بین داروها و بیماری‌ها است. روش‌های مبتنی بر اتصال، داروی مورد نیاز را را بر اساس ساختار سه بعدی پروتئین‌ها پیش‌بینی می‌کند. این روش زمانی که ساختار پروتئین ناشناخته باشد، کارایی خود را از دست می‌دهد. رویکردهای مبتنی بر شبکه و یادگیری ماشین تلاش می‌کنند تا بر محدودیت‌های ذکر شده در دو رویکرد دیگر غلبه کنند. این دو رویکرد نیز در پیش‌بینی موثر داروها به عواملی مانند داده‌های از پیش تعریف شده و قابل اعتماد بستگی دارند (۱۶).

**برهمکنش پروتئین-پروتئین و درمان بیماری‌ها:** در سیستم بدن موجودات زنده پروتئین‌ها با یکدیگر تعاملاتی را برقرار می‌کنند که در نتیجه آن بدن قادر می‌شود تا فعالیت‌های روزمره خود را انجام دهد. چنین تعاملاتی برهمکنش پروتئین-پروتئین نامیده می‌شوند. برهمکنش پروتئین-پروتئین در درمان بیماری‌ها و ایجاد اطلاعات اولیه برای کشف داروها تاثیر بسیار مهمی دارد. به عنوان مثال پروتئین p53 در بدن انسان یک پروتئین سرکوبگر تومور است که نقش حیاتی در حفظ پایداری ژنوم و جلوگیری از رشد تومورهای سرطانی، برای نمونه در سرطان ریه، را ایفا می‌کند (۱۷). در حدود ۵۰ درصد از سرطان‌های انسانی، p53 جهش می‌یابد و در نتیجه، کاهش سطح این پروتئین در بسیاری از سرطان‌ها مشاهده می‌گردد. داروهای رایج برای درمان بیماری‌های سرطانی به راحتی نمی‌توانند بر روی پروتئین p53 تاثیر بگذارند و چالش کشف دارو برای کنترل این پروتئین در درمان سرطان بسیار مهم است (۱۸). با بررسی شبکه برهمکنش پروتئین- پروتئین در بدن انسان مشخص شده است که پروتئین p53 با دو پروتئین MDM2 و MDM4 برهمکنش دارد و این دو پروتئین فعالیت P53 را کنترل می‌کنند. MDM2 باعث بی‌ثباتی و تجزیه سریع P53 می‌گردد. همچنین پروتئین MDM4 نیز از فعالیت این پروتئین می‌کاهد. دانشمندان با توجه به این اطلاعات به دنبال کشف گروهی از داروها بودند تا بتوان برهمکنش میان این پروتئین‌ها بایگدیگر را متوقف کنند که سرانجام با کشف گروهی از داروها به نام Nutlins، با موفقیت از اتصال MDM2 به پروتئین P53 جلوگیری شد. با مهار این اتصال، پروتئین P53 با ثبات‌تر شده و فعالیت سرکوب تومورهای سرطانی افزایش یافته است (۱۹). بنابر آنچه گفته شد پیش‌بینی برهمکنش‌های پروتئین- پروتئین در درمان بیماری‌ها و کشف دارو نقشی ضروری دارد. بسیاری از برهمکنش‌های موجود بین پروتئین‌ها همچنان ناشناخته باقی مانده است و بررسی وجود و یا عدم وجود این روابط در علوم تجربی و آزمایشگاه‌ها پرهزینه، زمان‌بر و دارای خطا می‌باشد. با استفاده از نظریه گراف، پردازش سیگنال، مدل‌های احتمالی و یادگیری ماشین می‌توان در علم داده به پیش‌بینی برهمکنش‌های پروتئین- پروتئین در سیستم موجودات زنده پرداخت. چنین روشی می‌تواند در مدت زمان کمتر و با دقتی مناسب برهمکنش‌های پروتئین-پروتئین را محاسبه کند.

**پردازش سیگنال گراف:** برهمکنش‌های موجود میان پروتئین‌ها را می‌توان با در نظر گرفتن شبکه‌ای بر روی آن‌ها مطالعه کرد. در چنین رویکردی هر پروتئین و برهمکنش‌های آن با دیگر پروتئین‌ها توسط گراف مدل سازی می‌شود؛ به‌طوریکه هر پروتئین به صورت یک راس و برهمکنش بین دو پروتئین دلخواه با یال نمایش داده می‌شود. فرض کنید گراف $G$ به صورت $\mathrm{G} = (\mathrm{V}, \mathrm{E})$ نمایش داده شده باشد که در آن $V$ مجموعه‌ای از $N$ راس $\{v_1, v_2, \dots, v_n\}$ متناظر با مجموعه پروتئین‌ها بوده و $E$ مجموعه برهمکنش‌های موجود در شبکه باشد به‌طوریکه $\forall v_i, v_j \in V : \{e_{ij} = (v_i, v_j)\}$ (۲۰). برای مثال در شکل ۱ نمونه‌ای از شبکه برهمکنش پروتئین-پروتئین بدن انسان توسط گراف مدل‌سازی شده است. این شبکه شامل ۸۱۴۹ پروتئین و ۱۱۵۲۳ برهمکنش می باشد. در گراف متناظر با این شبکه هر راس متناظر با یک پروتئین بوده و تمامی

راس‌ها با رنگ مشکی نمایش داده شده‌اند، همچنین در این گراف هر یال، متناظر با یک برهمکنش بین دو پروتئین بوده و تمامی یال‌ها با رنگ آبی مشخص شده‌اند.



شکل ۱- گراف برهمکنش پروتئین-پروتئین برای مجموعه داده شبکه پروتئین‌های انسان.

با استفاده از گراف $G$ می‌توان ماتریس مجاورت A را تعریف نمود که در آن مولفه $a_{ij}$ نشان‌دهنده وجود و یا عدم وجود یال بین دو راس i و j می‌باشد. لاپلاسین گراف $G$ به صورت $L = D - A$ تعریف می‌گردد که در آن $D$ ماتریس قطری درجه راس‌های گراف $G$ می‌باشد و عناصر این ماتریس به‌صورت $d_{ii} = \sum_{j=1}^{N} a_{ij}$ به ازای هر اندیس $i = 1, \dots, N$ تعریف می‌شوند. مقادیر ویژه ماتریس لاپلاسین را با $\lambda_1, \dots, \lambda_N$ نمایش می‌دهیم، مقادیر ویژه لاپلاسین گراف نقشی مشابه با فرکانس در تبدیل فوریه معمولی را ایفا می‌کنند. بردارهای ویژه متناظر با مقادیر ویژه، تشکیل ماتریس $U = [u_1, u_2, \dots, u_n]$ را می‌دهند. سیگنال گراف یک بعدی $x_n \in R$ برای راس دلخواه $n$ به صورت یک عدد حقیقی و سیگنال گراف چند بعدی $x_n \in R^M$ برای راس دلخواه $n$ به صورت یک بردار قابل تعریف می‌باشد.

**تبدیل موجک طیفی گراف:** اتصالات موضعی در گراف را می‌توان با استفاده از تبدیل موجک طیفی گراف ثبت نمود. در پردازش سیگنال کلاسیک، یک موجک خانواده‌ای از سیگنال‌های بومی‌سازی شده در زمان را فراهم می‌کند که برای نمایش ویژگی‌های سیگنال موضعی مورد استفاده قرار گرفته می‌شوند. بنابراین در پردازش سیگنال کلاسیک، عنصر پایه موجک، نه بر اساس مقیاس یا همان سرعت تغییرات آن، بلکه بر اساس مکان سیگنال پارامتر می‌شود (۲۱). در پردازش سیگنال گراف، عنصر پایه موجک با پارامتر گسسته n و پارامتر مقیاس پیوسته s مشخص می‌شود. پارامتر n موضعی سازی اطراف راس n را بررسی کرده و پارامتر s میزان هموار بودن تابع موجک را روی گراف کنترل می‌کند. با استفاده از پارامتر $s$ می‌توان یک تابع هسته باند گذر را تعریف نمود، به طوریکه این تابع بر روی مقادیر ویژه ماتریس لاپلاسین حرکت می‌کند. تابع هسته باند گذر $g$ به صورت زیر تعریف می‌گردد:

$$g(s\lambda) = \begin{cases} x_1^{-\alpha}(s\lambda)^{\alpha}, & s\lambda < x_1 \\ p\,(s\lambda), & x_1 \le s\lambda \le x\_2 \\ x_2^{\beta}(s\lambda)^{-\beta}, & s\lambda > x\_2 \end{cases}$$

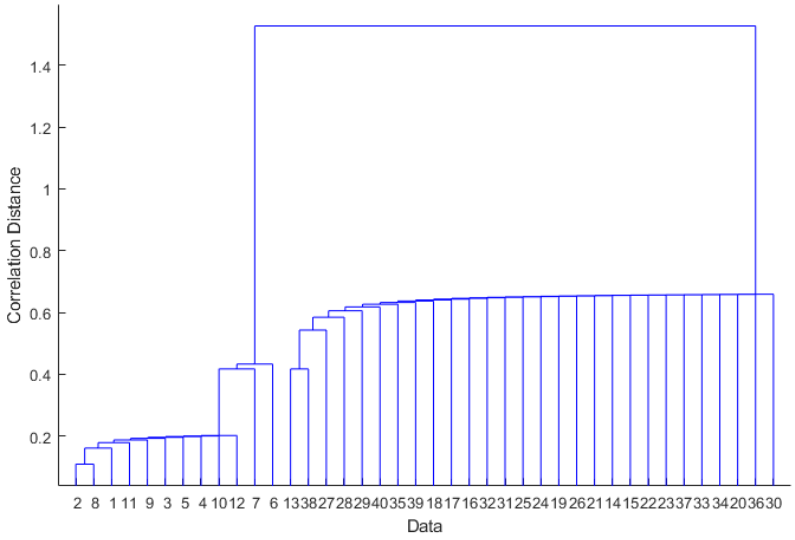
که در آن $p(s\lambda)$ به صورت یک درون‌یاب چندجمله‌ای مکعبی منحصر به فرد در نظر گرفته می‌شود که به پیوستگی g و مشتق آن توجه دارد. مقادیر $\alpha$ و $\beta$ و نقاط انتقال $x_1$ و $x_2$ پارامترهای فیلتر می‌باشند. در مساله پردازش شبکه برهمکنش پروتئین-پروتئین مقادیر ویژه $\lambda_1$ و $\lambda_2$ متناظر با طیف‌های بیشترین درجات تمامی راس‌ها در گراف می‌باشند و برای اطمینان از وجود آن‌ها در بررسی تبدیل موجک طیفی هر راس، $\alpha = \beta = \frac{1}{2}$ و $x_1 = 1$ و $x_2 = 2$ در نظر گرفته می‌شود. عنصر پایه موجک $\psi_{s,n}$ در مقیاس $s$ و حول راس $n$ به صورت زیر تعریف می‌گردد:

$$\psi_{s,n} = \sum_{k=1}^{n} g(s\lambda_n) u^{(k)}\,[n] u^{(k)}$$

تبدیل موجک طیفی گراف در حقیقت یک ترکیب خطی از بردارهای ویژه لاپلاسین گراف $G$ می‌باشد. اطلاعات موضعی و توپولوژیکی راس‌های گراف در تبدیل موجک طیفی گراف ثبت می‌گردد، بنابراین مقدار $\psi_{s,n}$ ویژگی راس $n$ در مقیاس $s$ نامیده می‌شود. با استفاده از ویژگی محاسبه شده برای هر راس، فاصله بین دو راس دلخواه $i$ و $j$ در مقیاس $s$ به صورت زیر تعریف می‌گردد:

$$D_s(i,j) = 1 - \frac{\psi_{s,i}^T \psi_{s,j}}{\|\psi_{s,i}\| \|\psi_{s,j}\|}$$

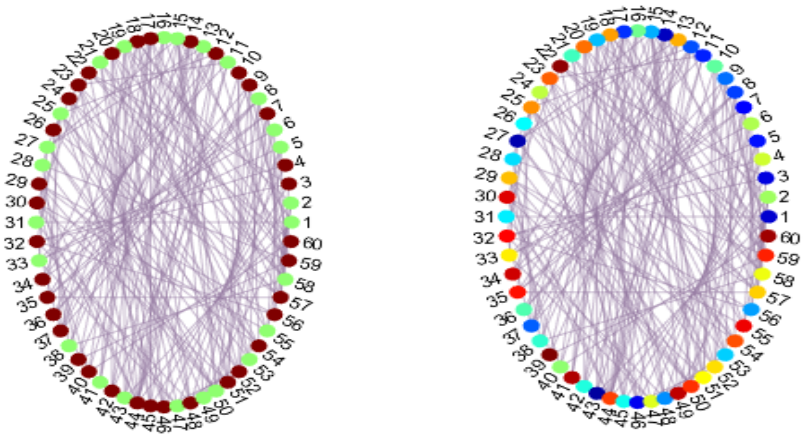
**خوشه‌بندی داده‌ها**: برای گروه‌بندی داده‌های موجود در یک مجموعه می‌توان از خوشه‌بندی استفاده نمود. خوشه‌بندی فرایند یافتن دسته‌هایی از داده‌ها می‌باشد که عناصر موجود در هر دسته، دارای ویژگی‌های یکسانی نسبت به یکدیگر هستند. در فرایند خوشه‌بندی مجموعه‌ای از داده‌ها بر اساس شباهت و یا عدم شباهت در زیرمجموعه‌هایی مختلف به نام خوشه، طبقه‌بندی می‌شوند. خوشه‌بندی سلسله مراتبی یک روش محبوب برای دسته‌بندی داده‌های تعریف شده بر روی گراف، از جمله شبکه‌های برهمکنش پروتئین-پروتئین است. در این روش یک داده ممکن است در بیش از یک خوشه حضور داشته باشد. خوشه‌بندی سلسله مراتبی را می‌توان با استفاده از الگوریتم‌های مختلفی مانند الگوریتم پیوند تکی، پیوند کامل و پیوند میانگین اجرا نمود. در الگوریتم پیوند میانگین، فاصله بین دو خوشه دلخواه $C_i$ و $C_j$ به صورت میانگین فاصله تمامی داده‌های موجود در دو خوشه تعریف می‌گردد. با اجرای خوشه‌بندی سلسله مراتبی با الگوریتم پیوند میانگین در مقیاس دلخواه $s$، راس‌های گراف $G$ به $K$ خوشه $\Gamma_k^{(s)}$ تقسیم می‌گردند که در آن $k = 1, \dots, K$ (۲۲). به‌عنوان مثال در شکل ۲ مجموعه ۴۰ پروتئین از بدن کرم الگانس با استفاده از این روش خوشه‌بندی شده است.




شکل ۲- خوشه‌بندی ۴۰ پروتئین از مجموعه پروتئین‌های بدن کرم الگانس با استفاده از الگوریتم پیوند میانگین.

با در نظر گرفتن یک حد آستانه برای فاصله همبستگی در نمودار خوشه‌بندی داده‌ها، نمودار قطع شده و داده‌ها به تعدادی خوشه تقسیم می‌گردند. حد آستانه به صورت میانگین فاصله همبستگی تمامی خوشه‌های بوجود آمده در نظر گرفته می‌شود. به عنوان مثال در شکل ۳، با محاسبه تبدیل موجک طیفی گراف ۶۰ راس از گراف متناظر با شبکه برهمکنش پروتئین-پروتئین بدن انسان، خوشه‌بندی سلسله مراتبی با الگوریتم پیوند میانگین در مقیاس $S = 1,2$ اجرا شده است، راس‌های قرار گرفته در یک خوشه یکسان، با رنگ مشابه نمایش داده شده‌اند.

**میدان تصادفی مارکوف**: میدان تصادفی مارکوف در نظریه آمار و احتمال برای توصیف تعاملات شرطی داده‌های همسایه در یک شبکه، مانند شبکه برهمکنش پروتئین‌ها استفاده می‌گردد. میدان تصادفی یک نوع فرایند تصادفی یک بعدی است که می‌تواند متغیرهای تصادفی را برحسب موقعیت مکانی آن‌ها مدل سازی کند. به عبارت دیگر در این فرایند داده‌ها برحسب داده‌های همسایه خود مدل می‌شوند (۵).



(الف) $s = 1$ (ب) $s = 2$

شکل ۳- خوشه‌بندی سلسله مراتبی ۶۰ راس از گراف متناظر با شبکه برهمکنش پروتئین-پروتئین بدن انسان به ازای مقیاس‌های $S = 1,2$.

فرض کنید به ازای هر راس $v \in V$ از گراف $G$، سیگنال گراف $m$ بعدی $x(v) \in \mathbb{R}^m$ داده شده باشد. همچنین فرض کنید مجموعه راس‌های گراف $G$ به $C$ خوشه مجزا تقسیم شده باشد، در اینصورت برای هر راس دلخواه $v$ از گراف $G$، به ازای هر خوشه $c \in \{1,2,\dots,C\}$ می‌توان بردار $\gamma(v) = [\gamma_1(v), \gamma_2(v), \dots, \gamma_C(v)]$ را تعریف نمود، به‌طوریکه مولفه $\gamma_c(v) = 1$ اگر و تنها اگر راس $v$ به خوشه $c$ متعلق باشد و در غیر اینصورت مقدار صفر را داشته باشد. به ازای هر راس $v \in V$، مجموعه سیگنال‌های گراف $x(v)$ و بردارهای $\gamma(v)$ تشکیل میدان‌های تصادفی $X: V \to \mathbb{R}^M$ و $\Gamma: V \to B$ را می‌دهند که در آن $B$ مجموعه بردارهای $\gamma(v)$ به ازای تمامی راس‌های گراف $G$ می‌باشد. همچنین به ازای هر یال $e$ از گراف $G$ مجموعه مولفه‌های ماتریس مجاورت $A$ نیز تشکیل میدان تصادفی مارکوف را می‌دهند. می‌توان اثبات نمود که با توجه به سه میدان نام برده فضای $M \triangleq V \cup E \cup V$ موجود بوده و یک میدان تصادفی مارکوف به صورت $\tilde{X}: M \to \mathbb{R}^M \cup \{0,1\} \cup B$ قابل تعریف می‌باشد، به‌طوریکه در آن به ازای هر $m \in [v,e,v]$ داریم:

$$\tilde{x}(m) = \begin{cases} x_v \in \mathbb{R}^m \\ a_e \in \{0,1\} \\ \gamma_v \in B \end{cases}$$

چنین میدانی در توزیع گیبس (۳) صدق می‌کند و لذا احتمال وجود یک عنصر در میدان مارکوف از رابطه زیر بدست می‌آید:

$$P_{\tilde{X}}(\tilde{x}) = \frac{exp\{-U(\tilde{x})\}}{\sum_{\tilde{x} \in \tilde{X}} exp\{-U(\tilde{x})\}}$$

که در آن U تابع پتانسیل سراسری بوده و به صورت زیر قابل محاسبه می‌باشد:

$$U(\tilde{x}) = \sum_{m \in \tilde{X}} \sum_{n \in N_m} Q(\tilde{x}_m, \tilde{x}_n)$$

$Q$ تابع پتانسیل دسته‌ای بین دو عنصر $m$ و $n$ بوده و به صورت یک تابع غیرکاهشی نامنفی تعریف می‌گردد. $N_m$ مجموعه همسایگی‌های عنصر $m$ می‌باشد. با تعریف مناسب تابع پتانسیل دسته‌ای می‌توان احتمال وجود یک یال را در گراف $G$ بررسی نمود، به‌طوریکه این احتمال بر حسب مجموعه سیگنال‌های گراف و خوشه‌ها بدست می‌آید. در میدان تصادفی مارکوف $\tilde{X}$، احتمال توزیع گیبس وابسته به سه مولفه $X$، $A$ و $\Gamma$ می‌باشد بنابراین داریم:

$$P_{X,A,\Gamma}(\mathrm{x}, \mathrm{a}, \gamma) = \frac{exp\{-U(\mathrm{x}, \mathrm{a}, \gamma)\}}{Z}$$

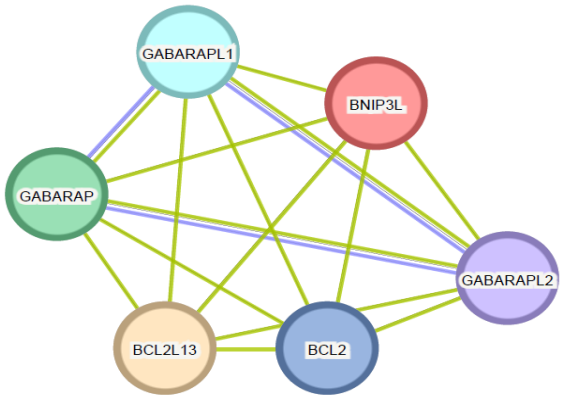
که در آن $Z$ یک عامل نرمال ساز برای داده‌ها بوده و از رابطه $Z = \sum_{\tilde{x} \in \tilde{X}} exp\{-U(\tilde{x})\}$ قابل محاسبه است. با توجه به سه مولفه $X$، $A$ و $\Gamma$ تابع پتانسیل سراسری را می‌توان از رابطه صفحه بعد محاسبه نمود:

$$U(\mathrm{x},\mathrm{a},\gamma)=\sum_{i\in v,\ j\in N_i} a_{ij}Q_X(x_i,x_j)+\sum_{i\in v,\ j\in N_i} a_{ij}Q_\Gamma(\gamma_i,\gamma_j)+\sum_{e\in E,\ k\in N_e} Q_A(a,a_k)$$

$Q_X$، $Q_\Gamma$ و $Q_A$ سه تابع پتانسیل دسته‌ای و نامنفی می باشند. $\sum_{i\in v,j\in N_i} a_{ij}Q_X(x_i,x_j)$ مرتبط با سیگنال گراف بوده و میزان ارتباطات میان راس‌ها را در نظر می‌گیرد، $\sum_{i\in v,j\in N_i} a_{ij}Q_\Gamma(\gamma_i,\gamma_j)$ احتمال مشاهده یال‌ها را در خوشه‌های مختلف کنترل کرده و $\sum_{e\in E,k\in N_e} Q_A(a,a_k)$ یال‌های خارج شده از راس مفروض را در خود ذخیره می‌کند.

**طرح الگوریتم پیش‌بینی برهمکنش پروتئین-پروتئین:** در این بخش به بررسی یک الگوریتم برای حل مساله پیش‌بینی برهمکنش پروتئین-پروتئین با استفاده از مدل گراف می‌پردازیم. در شکل ۴، مساله پیش‌بینی یال‌های ناشناس در بخشی از شبکه پروتئین درون سلولی نشان داده شده است. پروتئین‌های GABARAPL1، BNIP3L، GABARAPL2، BCL2، BCL2L13 و GABARAP در سلول با حذف میتوکندری تا سطح پایه به تنظیم کمیت و کیفیت میتوکندری کمک می‌کنند. در مساله پیش‌بینی برهمکنش پروتئین-پروتئین در این شبکه، برهمکنش‌های موجود بین پروتئین‌های GABARAPL1، GABARAPL2 و GABARAP حذف گردیده و با استفاده از الگوی اتصال دیگر پروتئین‌ها، احتمال وجود برهمکنش بین پروتئین های ذکر شده بررسی می‌شود. در شکل ۴ یال‌های سبز متناظر با برهمکنش‌های اولیه می‌باشد. الگوریتم بر روی داده‌ها اجرا شده و یال‌های آبی رنگ را در مجموعه به عنوان برهمکنش، پیش‌بینی نموده است. الگوریتم پیشنهادی شامل سه مرحله می‌باشد.

**مرحله اول،** گروه‌بندی داده‌ها با استفاده از تبدیل موجک طیفی گراف: در ابتدا با استفاده از تبدیل موجک طیفی گراف، ویژگی‌های متناظر با هر راس دلخواه $i$ در مقیاس s به صورت $\psi_{s,i}$ استخراج می‌شود. با استفاده از ویژگی‌های بدست آمده برای تمامی راس‌های موجود، فاصله همبستگی و خوشه‌بندی سلسله مراتبی با پیوند میانگین می‌توان داده‌ها را به $K$ خوشه $\Gamma_k^{(s)}$ گروه‌بندی نمود، به‌طوریکه در آن $k=1,\dots,K$. هر خوشه متناظر با زیرمجموعه‌ای از پروتئین‌ها و برهمکنش‌های موجود میان آن‌ها است.




شکل ۴- پیش‌بینی یال ناشناس در شبکه گراف برهمکنش پروتئین-پروتئین بدن انسان، یال‌های شناخته شده با رنگ سبز و یال ناشناس محاسبه شده توسط الگوریتم با رنگ آبی مشخص شده‌اند.

**مرحله دوم،** تعریف سیگنال گراف چندبعدی: حال به تعریف سیگنال گراف چندبعدی برای هر راس قرار گرفته در هر خوشه می‌پردازیم. به ازای هر راس دلخواه $i$ سیگنال گراف چندبعدی برای آن به صورت زیر قابل تعریف می باشد:

$$x_i = A_i$$

که در آن $A_i$ ستون $i$-ام از ماتریس مجاورت گراف می‌باشد. با توجه به تعریف سیگنال گراف چندبعدی معرفی شده در تابع پتانسیل سراسری می‌توان از مولفه $\sum_{e\in E,k\in N_e} Q_A(a,a_k)$ صرف نظر نمود زیرا در این مدل از سیگنال گراف چند بعدی، تمامی یال‌های خارج شده از تمامی راس‌ها در نظر گرفته می‌شود.

**مرحله سوم**، معرفی یک نگاشت برای تخمین سیگنال گراف مارکوف: همان‌طور که از احتمال وجود یک عنصر در میدان مارکوف مشخص است احتمال وجود یال بین دو راس $i$ و $j$ زمانی در بیشترین حالت ممکن خود قرار خواهد داشت که تابع پتانسیل کمترین مقدار ممکن را به خود بگیرد و لذا برای حل مساله پیش‌بینی برهمکنش پروتئین-پروتئین به‌ازای هر یال ناشناس $\bar{a}_{ij}$ بایستی تساوی زیر برقرار باشد:

$$\bar{a}_{ij} = \min_{a_{ij} \in A} U(x, \gamma)$$

پروتئین‌ها زمانی بیشترین احتمال برهمکنش را بایکدیگر دارند که در گراف متناظر با شبکه برهمکنش داده‌ها بین دو پروتئین مسیری با سه راس موجود باشد (۲۳)، ولذا تابع پتانسیل دسته‌ای بایستی به گونه‌ای تعریف گردد که در این شرط صدق کند. بدین منظور تابع پتانسیل دسته‌ای به صورت زیر تعریف می‌گردد:

$$Q_X(x_i, x_j) = \mu_\chi\left(1 - \frac{\|X_i A X_j\|_{1,1}}{\|A\|_{1,1}}\right)$$

که در آن $0 \le \mu_\chi \le 1$ در نظرگرفته می‌شود. تابع پتانسیل بین دسته‌ای نیز از رابطه زیر قابل محاسبه می‌باشد:

$$Q_\Gamma(\gamma_i, \gamma_j) = \|\gamma_i - \gamma_j\|_0$$

بنابر آنچه گفته شد مساله پیش‌بینی یال در گراف $G$ را می‌توان به صورت زیر بازنویسی نمود:

$$\bar{a}_{ij} = \min_{a_{ij} \in A} \sum_{i \in v, j \in N_i} \bar{a}_{ij} Q_X(x_i, x_j) + \sum_{i \in v, j \in N_i} \bar{a}_{ij} Q_\Gamma(\gamma_i, \gamma_j)$$

همان‌طور که در رابطه فوق مشخص است مدل میدان تصادفی مارکوف و نحوه تعریف تابع پتانسیل بر روی سیگنال گراف نقش مهمی را در حل مساله ایفا می‌کند. هر چه تابع پتانسیل الگوی بهتر و واقعی‌تری را از شبکه برهمکنش داده‌ها دریافت کند، مدل میدان تصادفی مارکوف با استفاده از توزیع گیبس احتمال دقیق‌تری از وجود برهمکنش در میان داده‌ها را تشخیص خواهد داد. با در نظر گرفتن حد آستانه $\theta$ می‌توان مساله پیش‌بینی برهمکنش پروتئین-پروتئین را حل نمود. اگر $\bar{a}_{ij}$ از حد آستانه $\theta$ بزرگتر باشد یک یال در گراف متناظر با مجموعه داده بین دو راس $i$ و $j$ رسم گردیده و در نظر گرفته می‌شود که دو پروتئین $i$ و $j$ با یکدیگر برهمکنش دارند. حد آستانه $\theta = 0.95$ در نظر گرفته شده است.

**الگوریتم ۱- پیش‌بینی برهمکنش پروتئین- پروتئین بر مبنای گراف (GRABP) (۲).**

- دریافت ماتریس مجاورت $A$، حد آستانه $\theta$.
- خوشه‌بندی گراف متناظر با شبکه برهمکنش پروتئین- پروتئین به $K$ خوشه با استفاده از خوشه‌بندی سلسله مراتبی با الگوریتم پیوند میانگین.
- محاسبه سیگنال گراف $x_i$ و $x_j$ برای راس $i$ و $j$ از گراف $G$ با استفاده از ماتریس مجاورت $A$.
- حل مساله پیش‌بینی یال در گراف $G$ از رابطه:

$$\bar{a}_{ij} = \min_{a_{ij} \in A} \sum_{i \in V, j \in N_i} \bar{a}_{ij} Q_X(x_i, x_j) + \sum_{i \in V, j \in N_i} \bar{a}_{ij} Q_\Gamma(\gamma_i, \gamma_j)$$

- اگر $\bar{a}_{ij} \ge \theta$ در گراف متناظر با شبکه برهمکنش پروتئین- پروتئین بین دو راس $i$ و $j$ یالی در نظر گرفته بشود.

**پیاده سازی الگوریتم پیشنهادی و ارزیابی عملکرد:** در این بخش الگوریتم پیش‌بینی برهمکنش پروتئین-پروتئین با استفاده از پردازش سیگنال گراف را بر روی مجموعه داده‌های واقعی اعمال می‌کنیم. مجموعه این آزمایش‌ها در نرم‌افزار متلب با درنظر گرفتن ۱۰ تکرار مونته کارلو انجام شده است. در هر تکرار تعدادی یال به‌صورت تصادفی از گراف متناظر با شبکه برهمکنش پروتئین- پروتئین حذف شده و مساله پیش‌بینی برهمکنش داده‌ها اعمال شده است. در این مقاله از مجموعه داده‌های برهمکنش‌های پروتئین-پروتئین بدن انسان (۲۴)، گیاه رشادی (۲۵)، کرم الگانس (۲۶) و مخمر (۲۷) استفاده شده است. تعداد پروتئین‌ها، تعداد برهمکنش‌ها و تعداد یال‌های حذف شده در هر مرحله برای هر مجموعه داده معرفی شده، در جدول ۱ ارائه شده است.

جدول ۱- تعداد پروتئین‌ها، تعداد برهمکنش‌ها و تعداد یال‌های حذف شده در اجرای الگوریتم.

| مجموعه داده | تعداد راس‌ها | تعداد یال‌ها | تعداد یال‌های حذف شده |
|---|---|---|---|
| بدن انسان | ۸۱۴۹ | ۱۱۵۲۳ | ۱۱۵۲ |
| گیاه رشادی | ۲۵۳۲ | ۵۹۱۹ | ۵۹۱ |
| کرم الگانس | ۲۲۱۴ | ۳۵۳۸ | ۳۵۳ |
| مخمر | ۱۶۴۷ | ۲۵۱۸ | ۲۵۱ |

در تمامی مجموعه داده‌ها $\mu_\chi = 0.75$ و برای برقراری بهینگی الگوریتم و جلوگیری از هزینه محاسباتی $S_{max} = 6$ در نظر گرفته شده است. الگوریتم پیش‌بینی برهمکنش، خوشه‌بندی سلسله مراتبی با پیوند میانگین را در ۶ مقیاس مختلف اجرا کرده و در هر تکرار مونته کارلو بهترین مقیاس را برای پیش‌بینی یال‌ها در گراف متناظر با شبکه داده‌ها را در نظر می‌گیرد. یکی از معیارهای مهم ارزیابی مساله یادگیری ماشین، منحنی مشخصه عملکرد می‌باشد که درصد تشخیص مثبت حقیقی را بر مبنای درصد تشخیص مثبت کاذب برای آزمایش مورد نظر رسم می کند (۲۸). استفاده از منحنی مشخصه عملکرد در ارزیابی مدل‌های یادگیری ماشین به دلیل دقت بالا و قابلیت مقایسه پذیری آن، بسیار مفید است. با درنظر گرفتن مساحت زیر منحنی مشخصه عملکرد، می‌توان بررسی نمود که الگوریتم معرفی شده تا چه حد به درستی عمل نموده است. با استفاده از این معیار می‌توان مشخص نمود که برهمکنش‌های بدست آمده از این الگوریتم در مجموعه داده‌های واقعی تا چه حد درست تشخیص داده شده‌اند. مقادیر مساحت زیر منحنی مشخصه عملکرد بین ۰ تا ۱ قرار دارند و هرچه این مقدار بیشتر باشد، نشان‌دهنده بهتر بودن عملکرد مدل است. به‌طور کلی عملکرد الگوریتم پیش‌بینی برهمکنش پروتئین-پروتئین با استفاده از پردازش سیگنال گراف و زمان محاسبه الگوریتم به ازای هر راس از مجموعه داده‌ها در جدول ۲ آمده است. در مجموعه داده شبکه برهمکنش‌های پروتئین- پروتئین بدن انسان، الگوریتم موفق به کسب مساحت 0.92 از منحنی مشخصه عملکرد شده است که این نتیجه نشان می‌دهد تابع پتانسیل بدلیل در نظر گرفتن مسیرهای به طول دو الگوی دقیقی را از شبکه برهمکنش داده‌ها بدست آورده است. الگوریتم در مجموعه داده‌های گیاه رشادی، کرم الگانس و مخمر نیز موفق به کسب مساحت مناسبی از منحنی مشخصه عملکرد شده است. با اجرای الگوریتم به ازای مقیاس‌های مختلف مشاهده می‌شود که بهترین نتایج از مقیاس $s = 3$ حاصل می‌شوند زیرا در این مقیاس تابع هسته باند گذر $g$ از طیف مناسبی از مقادیر ویژه ساخته می‌شود و همچنین تبدیل موجک طیفی گراف یا همان ویژگی متناظر با هر راس نسبت به بالاترین ۳ درجه موجود در گراف محاسبه شده و بیشترین الگوهای اتصال در گراف را در نظر می‌گیرد.

جدول ۲- عملکرد الگوریتم پیش‌بینی برهمکنش پروتئین-پروتئین با استفاده از پردازش سیگنال گراف.

| مجموعه داده | مساحت زیر منحنی مشخصه عملکرد | زمان محاسبات به ازای هر راس گراف |
|---|---|---|
| بدن انسان | 0.92 | $2.20(s)$ |
| گیاه رشادی | 0.88 | $1.18(s)$ |
| کرم الگانس | 0.77 | $0.88(s)$ |
| مخمر | 0.76 | $0.70(s)$ |

**الگوریتم ارتباط طیفی**: الگوریتم ارتباط طیفی یا همان $spectral\ link$ یک نوع از خوشه‌بندی گراف بر پایه ماتریس لاپلاسین و الگوریتم خوشه‌بندی $K$-میانگین می‌باشد. فرض کنید $U = [u_1, u_2, \dots, u_n]$ ماتریس بردارهای ویژه متناظر با مقادیر ویژه ماتریس لاپلاسین گراف باشد. در این روش بردار ویژه راس $i$-ام از گراف $G$ به صورت ردیف $i$-ام از ماتریس $U$ تعریف شده و فاصله طیفی دو راس دلخواه $i$ و $j$ از رابطه:

$$D(i,j) = \|u_i - u_j\|$$

محاسبه می‌گردد که در آن به ترتیب $u_i$ و $u_j$ بردارهای ویژه متناظر با دو راس $i$ و $j$ از گراف $G$ می‌باشد. در این روش با استفاده از الگوریتم خوشه‌بندی $K$-میانگین، داده‌ها به $K$ خوشه مجزا دسته‌بندی شده و به هر راس دلخواه $i$ در خوشه p، $(p \in 1, \dots, K)$ مقدار $IDX(i) = \|u_i - c_p\|^2$ نسبت داده می‌شود که در آن $c_p$ مرکز خوشه $p$ می‌باشد. اگر دو راس دلخواه $i$ و $j$ در یک خوشه قرار گرفته باشند، میزان شباهت آن‌ها از رابطه:

$$Sim\ SC\ (i,j) = 1 - \left|D\big(i, IDX(i)\big) - D\big(j, IDX(j)\big)\right|$$

و اگر دو راس مفروض در یک خوشه قرار نگرفته باشند، میزان شباهت آن‌ها از رابطه:

$$Sim\ DC(i,j) = \frac{1}{1 + D\big(i, IDX(j)\big) + D\big(j, IDX(i)\big)}$$

محاسبه می‌گردد. اگر میزان شباهت دو راس $i$ و $j$ از حد آستانه $\theta$ بیشتر باشد این الگوریتم یالی بین دو راس $i$ و $j$ را بوجود آورده و به این صورت مساله پیش‌بینی برهمکنش پروتئین-پروتئین را در گراف متناظر با این شبکه حل می‌کند.

**الگوریتم ۲- پیش‌بینی برهمکنش پروتئین-پروتئین بر مبنای ارتباط طیفی (Spectral link) (۲۹).**

- دریافت ماتریس لاپلاسین گراف $G$، تعداد خوشه‌ها، حد آستانه $\theta$.
- محاسبه ماتریس بردارهای ویژه متناظر با مقادیر ویژه ماتریس لاپلاسین.
- محاسبه فاصله طیفی بین راس‌ها از رابطه $D(i,j) = \|u_i - u_j\|^2$.
- خوشه‌بندی راس‌ها به $K$ خوشه با استفاده از الگوریتم خوشه‌بندی $K$-میانگین.
- محاسبه میزان شباهت راس‌های دلخواه $i,j$ درون یک خوشه از رابطه:

$$Sim\ SC\ (i,j) = 1 - \left|D\big(i, IDX(i)\big) - D\big(j, IDX(j)\big)\right|$$

- محاسبه میزان شباهت راس‌های دلخواه $i,j$ موجود در خوشه‌های متفاوت از رابطه:

$$Sim\ DC(i,j) = \frac{1}{1 + D\big(i, IDX(j)\big) + D\big(j, IDX(i)\big)}$$

- بوجود آمدن یال به عنوان خروجی در صورتی که میزان شباهت از حد آستانه $\theta$ بیشتر باشد.

**الگوریتم پیش‌بینی برهمکنش پروتئین-پروتئین بر مبنای وجود مسیر ($L_3$):** این روش برای پیش‌بینی برهمکنش‌های موجود بین پروتئین‌ها، از الگوهای موجود در نحوه تعامل آن‌ها با یکدیگر استفاده می‌کند. برخلاف تصور رایج، شباهت ساختاری، تنها عامل برهمکنش پروتئین-پروتئین نمی‌باشد. الگوریتم $L_3$ با بررسی مسیرهای موجود بین پروتئین‌ها، به دنبال یافتن الگوهایی است که نشان‌دهنده احتمال برهمکنش می‌باشند. در این روش، اگر بین دو پروتئین، مسیری با طول سه وجود داشته باشد، الگوریتم پیش‌بینی می‌کند که این دو پروتئین با یکدیگر برهمکنش خواهند داشت. فرض کنید مولفه $a_{ij}$ از ماتریس مجاورت گراف، متناظر با وجود یال بین دو راس دلخواه $i$ و $j$ باشد. احتمال وجود یال بین این دو راس از رابطه:

$$p_{ij} = \frac{a_{iz}a_{zq}a_{qj}}{\sqrt{F_z F_q}}$$

محاسبه می‌گردد که در آن $F_z$ و $F_q$ به ترتیب درجه راس‌های $z$ و $q$ می‌باشند. اگر احتمال وجود برهمکنش بین دو راس $i$ و $j$ از حد آستانه $\theta$ بیشتر باشد، بین دو راس مفروض، یالی در نظر گرفته می‌شود.

**الگوریتم ۳- پیش‌بینی برهمکنش پروتئین-پروتئین بر مبنای وجود مسیر ($L_3$) (۲۳).**

- دریافت ماتریس مجاورت $A$، حد آستانه $\theta$.
- محاسبه احتمال وجود یال بین دو راس دلخواه $i$ و $j$ از رابطه:

$$p_{ij} = \frac{a_{iz}a_{zq}a_{qj}}{\sqrt{F_z F_q}}$$

- اگر $p_{ij} \geq \theta$، آنگاه بین دو راس $i$ و $j$ یالی در نظر گرفته می‌شود.

**مقایسه الگوریتم‌ها:** در این بخش به مقایسه سه الگوریتم پیش‌بینی برهمکنش پروتئین-پروتئین GRABP، Spectral link و $L_3$ می‌پردازیم. سه الگوریتم بررسی شده با در نظر گرفتن حد آستانه $\theta = 0.95$ بر روی سه مجموعه داده برهمکنش پروتئین-پروتئین بدن انسان، گیاه رشادی و کرم الگانس اجرا شده است و مساحت منحنی مشخصه عملکرد در جدول ۳ گزارش شده است.

جدول ۳- مقایسه منحنی مشخصه عملکرد در سه الگوریتم پیش‌بینی برهمکنش پروتئین-پروتئین بر مبنای گراف، ارتباط طیفی و وجود مسیر

| مجموعه داده | *GRABP* | *Spectral link* | $L_3$ |
|---|---|---|---|
| بدن انسان | 0.92 | 0.49 | 0.89 |
| گیاه رشادی | 0.88 | 0.66 | 0.82 |
| کرم الگانس | 0.77 | 0.50 | 0.70 |

همان‌طور که در جدول ۳ مشاهده می‌گردد، الگوریتم GRABP، قدرت پیش‌بینی بالاتری را از خود نشان می‌دهد. الگوریتم $L_3$، برهمکنش‌ها را به درستی تشخیص نمی‌دهد اما نسبت به الگوریتم Spectral link کارایی بالاتری را از خود نشان می‌دهد.

## نتیجه‌گیری

در این مقاله، پیش‌بینی برهمکنش پروتئین-پروتئین با استفاده از پردازش سیگنال گراف مرور شده است. بدین منظور ابتدا مدل شبکه برهمکنش پروتئین-پروتئین با

استفاده از گراف بررسی شده و سپس با استفاده از مفهوم تبدیل موجک طیفی گراف، ویژگی راس‌ها مطالعه شده است. در ادامه داده‌ها با استفاده از خوشه‌بندی سلسله مراتبی با الگوریتم پیوند میانگین گروه‌بندی شدند. در الگوریتم پیش‌بینی برهمکنش پروتئین-پروتئین بر مبنای گراف، مساله پیش‌بینی برهمکنش‌های ناشناخته با استفاده از میدان تصادفی مارکوف، به یک مساله بهینه‌سازی ماکسیمم احتمال وجود برهمکنش بین پروتئین‌ها تبدیل گردید. این مساله را به‌طور معادل می‌توان به‌صورت مساله کمینه شدن تابع پتانسیل مجموعه سیگنال‌ها و مجموعه خوشه‌ها در نظر گرفت. الگوریتم بررسی شده در ده تکرار بر روی گراف داده‌ها اجرا شد. با استفاده از معیار مساحت منحنی مشخصه عملکرد سه الگوریتم GRABP، Spectral link و $L_3$ مقایسه شد. نتایج نشان می‌دهند که الگوریتم پیش‌بینی برهمکنش پروتئین-پروتئین بر مبنای گراف (GRABP)، برای مجموعه داده‌های مختلف در موجودات زنده، توانایی بالاتری را نسبت به سایر روش‌ها در پیش‌بینی صحیح برهمکنش‌ها دارد.

## منابع


1- A. Ortega, P. Frossard, J. Kovačević, J. M. Moura, and P. Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5): 808–828, 2008.

2- S. Colonnese, M. Petti, L. Farina, G. Scarano, and F. Cuomo. Protein-protein interaction prediction via graph signal processing. *IEEE Acces*s, 9:142681–142692, 2021.

3- N. Tremblay and P. Borgnat. Graph wavelets for multiscale community mining. *IEEE Transactions on Signal Processing*, 62(20):5227–5239, 2014.

4- M. Cheung, J. Shi, O. Wright, L. Y. Jiang, X. Liu, and J. M. Moura. Graph signal processing and deep learning: Convolution, pooling, and topology. *IEEE Signal Processing Magazine*, 37(6): 139–149, 2020.

5- S. Colonnese, P. Di Lorenzo, T. Cattai, G. Scarano, and F. D. V. Fallani. A joint markov model for communities, connectivity and signals defined over graphs. *IEEE Signal Processing Letters*, 27:1160–1164, 2020.

6- M. Tsubaki, K. Tomii, and J. Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2): 309–318, 2019.

7- S. Vilar, E. Uriarte, L. Santana, T. Lorberbaum, G. Hripcsak, C. Friedman, and N. P. Tatonetti. Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nature Protocols*, 9(9):2147–2163, 2014.

8- X. Lin, Z. Quan, Z.-J. Wang, H. Huang, and X. Zeng. A novel molecular representation with BIGRU neural networks for learning atom. *Briefings in Bioinformatics*, 21(6):2099–2111, 2020.

9- T. Ma, C. Xiao, J. Zhou, and F. Wang. Drug similarity integration through attentive multi-view graph auto-encoders. International Joint Conference on Artificial Intelligence, 18:*3477-3483*, 2018.

10- B. Jin, H. Yang, C. Xiao, P. Zhang, X. Wei, and F. Wang. Multitask dyadic prediction and its application in prediction of adverse drug-drug interaction. *In Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

11- X. Chu, Y. Lin, Y. Wang, L. Wang, J. Wang, and J. Gao. Mlrda: A multi-task semi-supervised learning framework for drug-drug interaction prediction. *In Proceedings of the 28th international joint conference on artificial intelligence*, pages 4518–4524, 2019.

12- W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. *Advances in Neural Information processing systems*, 30, 2017.

13- P. Csermely, T. Korcsmáros, H. J. Kiss, G. London, and R. Nussinov. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology & Therapeutics*, 138(3):333–408, 2013.

14- J. Loscalzo and A.-L. Barabasi. Systems biology and the future of medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(6):619–627, 2011.

15- Gov, U. FDA drug development process, https://www.fda.gov/patients/learn-about-drug-

and-device-approvals/drug-development-process, Accessed on 4/10/2024.

16- L. Zhou, Z. Li, J. Yang, G. Tian, F. Liu, H. Wen, L. Peng, M. Chen, J. Xiang, and L. Peng. Revealing drug-target interactions with computational models and algorithms. *Molecules*, 24(9):1714, 2019.

17- D. R. Green and J. E. Chipuk. p53 and metabolism: Inside the tigar. *Cell*, 126(1):30–32, 2006.

18- F. Toledo and G. M. Wahl. Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nature Reviews Cancer*, 6(12):909–923, 2006.

19- L. T. Vassilev, B. T. Vu, B. Graves, D. Carvajal, F. Podlaski, Z. Filipovic, N. Kong, U. Kammlott, C. Lukacs, C. Klein, et al. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science*, 303(5659):844–848, 2004.

20- D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing magazine*, 30(3):83–98, 2013.

21- D. K. Hammond, P. Vandergheynst, and R. Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.

22- S. Zhou, Z. Xu, and F. Liu. Method for determining the optimal number of clusters based on agglomerative hierarchical clustering. *IEEE Transactions on Neural Networks and Learning systems*, 28(12):3007–3017, 2016.

23- I. A. Kovács, K. Luck, K. Spirohn, Y. Wang, C. Pollis, S. Schlabach, W. Bian, D. K. Kim, N. Kishore, T. Hao, et al. Network-based prediction of protein interactions. *Nature Communications*, 10(1):1240, 2019.

24- T. Rolland, M. Taşan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, et al. A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226, 2014.

25- H. Herrera-Ubaldo, S. E. Campos, P. López-Gómez, V. Luna-García, V. M. Zúñiga-Mayo, G. E. Armas-Caballero, K. L. González-Aguilera, A. DeLuna, N. Marsch-Martínez, C. Espinosa-Soto, et al. The protein–protein interaction landscape of transcription factors during gynoecium development in arabidopsis. *Molecular Plant*, 16(1):260–278, 2023.

26- S. Remmelzwaal and M. Boxem. Protein interactome mapping in caenorhabditis elegans. *Current Opinion in Systems Biology*, 13:1–9, 2019.

27- J. Chen and B. Yuan. Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290, 2006.

28- A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction networks. *Complexus*, 1(1):38–44, 2003.

29- P. Symeonidis and N. Mantas, “Spectral clustering for link prediction in social networks with positive and negative links,” *Social Network Analysis and Mining*, vol.3, pp.1433–1447, 2013.

# Applications of graph signal processing in predicting protein-protein interactions

**Babaei R. and Jamshidi A.A.***

**Advanced Systems Biology and Cancer Research Lab., School of Mathematics Statistics and Computer Science, College of Science, University of Tehran, Tehran, I.R. of Iran**

## Abstract

Protein-protein interactions are physical and chemical connections between two or more proteins that play a vital role in a wide range of cellular processes. Graph signal processing is a new and powerful tool for graph analysis, which has recently been successfully applied to protein-protein networks. In this article, the basic concepts of graph signal processing for protein-protein interaction network modeling are reviewed. Graph features are investigated using spectral graph wavelet transform and hierarchical clustering. Some of the main unknown interaction prediction algorithms including GRABP, Spectral link and $L_3$ algorithms are reviewed and the results of these algorithms are reported on protein-protein interaction graph networks of the human body, Arabidopsis plant, C. Elegans worm and yeast. The results of these algorithms show that graph signal processing can detect unknown interactions in the graph network of living organisms with good approximation.

**Key words:** Protein-protein interaction network, Graph signal processing, Spectral Graph wavelet transform, Machine learning.