دوره دهم، شماره اول، بهار و تابستان ۱۳۹۶

مجله فناوری اطلاعات در طراحی مهندسی
Information Technology in Engineering Design
http://ited.sinaweb.net

ساده‌سازی مدل هم‌ترازی شبکه‌های برهم‌کنش پروتئین- پروتئین با استفاده از ماتریس شباهت به منظور کاهش زمان حل

شادی طبسی‌کاخکی [۱] محمود نقیب‌زاده [۲] یحیی فرقانی* [۳]

(۱) گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

(۲) گروه مهندسی کامپیوتر، دانشگاه فردوسی مشهد، مشهد، ایران

(۳) گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران*

چکیده

مسئله هم‌ترازی شبکه‌های پروتئینی، یک مساله NP-Complete است. در این مقاله، برای کاهش پیچیدگی حل مدل ریاضی مساله هم ترازی، تقریبی از مدل ارائه می‌شود. به عبارت دقیق‌تر، از ماتریس شباهت دو شبکه پروتئینی برای کاهش تعداد متغیرهای مدل استفاده می‌گردد. برای این منظور، در مدل پیشنهادی، به‌جای بررسی هم‌ترازی هر یک از پروتئین‌های شبکه نخست با تمام پروتئین‌های شبکه دوم، هم‌ترازی هر یک از پروتئین‌های شبکه نخست فقط با تعدادی از شبیه‌ترین پروتئین‌های شبکه دوم بررسی می‌شود. مدل پیشنهادی، برای همترازی شبکه‌های پروتئینی واقعی گونه‌های مختلف و نیز شبکه‌های مصنوعی آزمایش شد. نتایج تجربی، نشان دهنده بهبود دقت هم‌ترازی نسبت به روش تقریبی NETAL و نیز کاهش زمان اجرا، نسبت به مدل دقیق می‌باشد. در ضمن، روش پیشنهادی، توانسته است بر روی شبکه‌های مصنوعی فاقد نویز، به دقت بسیار مطلوب دست یابد.

واژه‌های کلیدی: شبکه برهم‌کنش پروتئین-پروتئین، هم‌ترازی، ماتریس شباهت، مدلسازی ریاضی.

* عهده‌دار مکاتبات

نشانی: گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

تلفن: ۰۹۳۵۲۱۶۱۸۵۷ پست الکترونیکی: yforghani@mshdiau.ac.ir

۱- مقدمه

مطالعه در زمینه پروتئومیک به دلیل نقشی که پروتئین‌ها در فرآیندها و عملکردهای سلولی دارند روز به روز در حال گسترش است. از طرفی، عملکرد پروتئین‌ها در درون سلول از طریق برهم‌کنش با سایر پروتئین‌ها شکل می‌گیرد که این برهم‌کنش‌ها و تعاملات میان پروتئین‌ها، شبکه‌های برهم‌کنش پروتئین-پروتئین را شکل می‌دهد و مطالعه این شبکه‌ها در درک چگونگی عملکرد پروتئین‌ها در درون سلول حائز اهمیت است. در سال‌های اخیر، با به‌دست آمدن حجم زیادی از داده‌های شبکه‌های پروتئینی برای انسان و دیگر گونه‌های زیستی، روش‌های محاسباتی جهت مقایسه و تحلیل آن‌ها نیز گسترش یافته است. از این‌رو، در حوزه سیستم‌های زیستی، مقایسه معنادار شبکه‌های گونه‌های مختلف، مهم تلقی می‌شود و یکی از مهمترین روش‌های مقایسه میان شبکه‌ها، هم‌ترازی[1] آن‌ها می‌باشد [۱]. مسئله هم‌ترازی میان دو شبکه پروتئینی به‌طور خاص، یافتن بهترین تطابق از پروتئین‌های یک شبکه به پروتئین‌های شبکه دیگر است، به‌گونه‌ای که میان جفت پروتئین‌های هم‌تراز شده دو شبکه، بیشترین شباهت توپولوژیکی[2] و بیولوژیکی برقرار باشد. این امر منجر به شناسایی نواحی مشابه میان شبکه‌های گونه‌های مختلف می‌شود [۲]. از طرفی، هم‌ترازی این گونه شبکه‌ها رویکردی لازم و ضروری جهت درک عملکرد پروتئین‌های ناشناخته به شمار می‌رود [۳].

یکی از روش‌های حل مسئله هم‌ترازی، مدلسازی ریاضی و استفاده از روش‌های مختلف بهینه‌سازی برای حل مدل است. تاکنون چندین مدل جهت هم‌ترازی این شبکه‌ها پیشنهاد شده است [٤, ٥]. مساله هم‌ترازی شبکه‌های پروتئینی، یک مسئله NP-Complete است. لذا، الگوریتم کارایی برای حل مدل یادشده وجود ندارد [٦, ٧]. در مراجع [۲, ۸]، از روشهای فراابتکاری (تکاملی) برای حل این مدل استفاده شده است. شبکه‌های پروتئینی، دارای پروتئین‌های بسیار زیاد بوده و لذا هم‌ترازی دقیق آن‌ها با کامپیوترهای امروزی بسیار زمان‌بر یا ناممکن می‌باشد. برای رفع این مشکل، از روش‌های هم‌ترازی تقریبی استفاده می‌شود. در برخی از روش‌های تقریبی از جمله روش NETAL[۹]، از یک ماتریس شباهت برای حل تقریبی مساله هم‌ترازی استفاده می‌شود [۹-۱۱]. این روش ها، شامل دو فاز هستند. در فاز اول، شباهت هر جفت پروتئین از دو شبکه پروتئینی بر مبنای یک تابع هزینه گره یا NCF[3] محاسبه می‌گردد. ماتریس تشابه نشان می‌دهد که یک پروتئین از شبکه اول چقدر با پروتئین دیگر از شبکه دوم قابل انطباق است. در برخی از مراجع [11]، از ترکیب شباهت توپولوژیکی و بیولوژیکی برای ایجاد ماتریس شباهت استفاده می‌شود درحالیکه در برخی از مراجع [1,10]، تنها از معیارهای توپولوژیکی برای ایجاد ماتریس شباهت استفاده می‌شود. در فاز دوم این روش‌های هم ترازی تقریبی، از یک نوع استراژی همترازی یا AS[4] برای حل مساله هم‌ترازی با استفاده از ماتریس شباهت استفاده می‌شود [۷]. در بیشتر موارد، در این فاز، از رویکرد حریصانه "انتخاب و توسعه"[5] استفاده می‌شود [۱۰, ۱۲]. یعنی، ابتدا، جفت گره با بیش‌ترین امتیاز، هم‌تراز شده و سپس هم‌ترازی پیرامون آن گره با شروع از جفت گره‌هایی با بیشترین شباهت انجام می‌گیرد. روش مرجع [۱۱]، از ماتریس شباهت به عنوان وزن‌های گراف دو بخشی کامل استفاده کرده و مسئله حداکثر تطابق را با الگوریتم مجارستانی[6] حل می‌کند. در واقع، تفاوت روش‌های ابتکاری مبتنی بر ماتریس شباهت، در نحوه مدیریت کردن دو فاز یادشده است.

در مرجع [۵]، مدل ریاضی مساله هم‌ترازی، ساده شده است، یعنی مدل اولیه که یک مدل غیرخطی عدد صحیح است به یک مدل خطی عدد صحیح تبدیل شده است و برای حل مدل حاصل، از ساده سازی لاگرانژی[7] استفاده

[1] Alignment
[2] Topological

[3] Node Cost Function (NCF)
[4] Alignment Strategy (AS)
[5] Seed-and-Extend
[6] Hungarian
[7] Lagrangian relaxation

شده است. اگرچه حل مدل خطی عدد صحیح حاصل و الگوریتم تقریبی ارائه شده برای آن، بسیار کارآمدتر از الگوریتمهای عمومی حل مدل غیرخطی عدد صحیح اولیه است ولی هنوز هم اجرای این الگوریتم تقریبی به ازای شبکه‌های پروتئینی بزرگ در کامپیوترهای امروزی بسیار زمانبر یا ناممکن است و لذا، از یک کلاستر برای اجرای این الگوریتم استفاده شده است. در این مقاله نیز، برای کاهش پیچیدگی حل مدل ریاضی مساله هم ترازی، تقریبی از مدل ارائه می‌شود. به عبارت دقیق‌تر، از ماتریس شباهت دو شبکه پروتئینی برای کاهش تعداد متغیرهای مدل استفاده می‌گردد. برای این منظور، در مدل تقریبی پیشنهادی، به‌جای بررسی هم‌ترازی هر یک از پروتئین‌های شبکه نخست با تمام پروتئین‌های شبکه دوم، هم‌ترازی هر یک از پروتئین‌های شبکه نخست فقط با شبیه‌ترین پروتئین‌های شبکه دوم بررسی می‌شود. مدل پیشنهادی، برای همترازی شبکه‌های پروتئینی واقعی گونه‌های مختلف و نیز شبکه‌های مصنوعی آزمایش شد. نتایج تجربی، نشان دهنده بهبود دقت هم‌ترازی نسبت به روش تقریبی NETAL و نیز کاهش زمان اجرا، نسبت به مدل دقیق می‌باشد. در ضمن، روش پیشنهادی، توانسته است بر روی شبکه‌های مصنوعی فاقد نویز، به دقت بسیار مطلوب دست یابد. در ادامه، در بخش ۲، پیش نیازهایی از تحقیق و همچنین مدل ریاضی مساله هم‌ترازی شبکه‌های پروتئینی بیان می‌شود. در بخش ۳، مدل تقریبی پیشنهادی ارائه می‌گردد. در بخش ٤، نتایج آزمایش‌ها بر روی داده‌های مصنوعی و واقعی ارائه می‌شود و در بخش ۵، بحث و نتیجه گیری خواهد شد.

۲- پیش نیازها

۲-۱- شبکه‌های پروتئینی

پروتئین‌ها به‌تنهایی در سلول عمل نمی‌کنند و عملکرد آن‌ها به‌صورت برهم‌کنش با یکدیگر است. این برهم‌کنش‌ها، شبکه‌های برهم‌کنش پروتئینی را شکل می‌دهند. هر شبکه پروتئینی توسط گراف ساده بدون وزن و بدون جهت مدل می‌شود. گره‌های گراف، همان پروتئین‌ها و یال‌های آن، برهم‌کنش‌های فیزیکی میان پروتئین‌ها است که با روش‌های آزمایشگاهی به‌دست می‌آید. دو گراف بدون وزن و بدون جهت به‌صورت $G_1=(V_1,E_1)$ و $G_2=(V_2,E_2)$ را درنظربگیرید به طوری که گراف اول کوچک‌تر یا مساوی گراف دوم است ($|V_1|\leq|V_2|$). مسئله هم‌ترازی، یافتن تابع هم‌ترازی یک‌به‌یک f به‌صورت $f:V_1\rightarrow V_2$ است که هر گره از گراف اول را تنها به یک گره از گراف دوم متناظر می‌کند به طوری‌که شباهت میان گراف‌های G_1 و G_2 تحت هم‌ترازی معین f حداکثر گردد.

۲-۲- مدل ریاضی مساله هم‌ترازی شبکه‌های پروتئینی

یک مدل ریاضی برای مساله هم‌ترازی شبکه‌های پروتئینی [۵, ۱۳] به صورت زیر می‌باشد:

(۱)

$$min_p||A_G - PA_HP^T||_F^2$$
$$subject\ to \begin{cases}\sum_i P_{ij} = 1, & \forall j\\ \sum_j P_{ij} = 1, & \forall i\\ P_{ij} \in \{0,1\}, & \forall i,j\end{cases}$$

این مدل، یک مدل غیرخطی عدد صحیح محسوب می‌شود. در این مدل، دو شبکه پروتئینی در قالب دو گراف ساده، بدون جهت و بدون وزن به‌صورت $G=(V_1,E_1)$ و $H=(V_2,E_2)$ درنظرگرفته شده‌اند. هر گراف را می‌توان با ماتریس مجاورت متقارن A با سایز $|V|\times|V|$ نمایش داد که $|V|$ تعداد رئوس آن گراف است. در این ماتریس، درایه سطر i ام از ستون j ام برابر با یک است ($A_{ij}=1$)، اگر یالی میان گره i و j در گراف مربوطه وجود داشته باشد، در غیر این صورت، این درایه برابر با صفر می‌باشد. در مدل (۱)، A_G ماتریس مجاورت گراف G، A_H ماتریس مجاورت گراف H و دو گراف G و H هم اندازه فرض می‌شوند. البته در مسائل هم‌ترازی واقعی، تعداد پروتئین‌ها در دو شبکه پروتئینی، متفاوت است. در این صورت بایستی سایز شبکه کوچک‌تر با شبکه بزرگ‌تر یکسان گردد تا بتوان حاصل تفاضل ماتریس‌ها را با تابع هدف مدل (۱) به‌دست آورد زیرا ماتریس جایگشت P هم اندازه با ماتریس بزرگ‌تر

می‌باشد. درایه‌های ماتریس جایگشت برابر با صفر یا یک هستند و مجموع درایه‌های هر سطر یا هر ستون از این ماتریس برابر با یک است. قیدهای مدل (۱)، ماتریس P را به نحوی انتخاب می‌کنند که دارای ویژگی‌های یک ماتریس جایگشت باشد. ضرب ماتریس جایگشت P از سمت چپ در ماتریس مجاورت A_H می‌تواند جای سطرهای ماتریس A_H را تغییر دهد و ضرب ترانهاده ماتریس جایگشت P از سمت راست در ماتریس مجاورت A_H می‌تواند جای ستون‌های ماتریس A_H را تغییر دهد. اگر P_{ij}=1 آنگاه سطر و ستون j–ام ماتریس مجاورت A_H به سطر و ستون i–ام منتقل می‌شود. تغییر سطرها و ستون‌های ماتریس مجاورت A_H معادل با تغییر شماره گره‌های گراف H است. در تابع هدف مدل (۱) پس از تغییر شماره گره‌های گراف H، اختلاف ماتریس مجاورت گراف H با شماره گره‌های جدید و ماتریس مجاورت گراف G با استفاده از نرم فروبنیوس[1] محاسبه می‌شود. به تعبیر دیگر، تابع هدف مدل (۲)، به دنبال تغییر شماره گره‌های گراف H است به نحوی که اختلاف ماتریس مجاورت گراف G و ماتریس مجاورت گراف H با شماره گره‌های جدیدش، حداقل گردد. اگر در جواب بهینه مدل (۱)، گره شماره j گراف H به گره شماره i تغییر یافته باشد یا در نقطه بهینه، P_{ij}=1 شده باشد آنگاه گره شماره j گراف H گره هم تراز گره شماره i گراف G است. درضمن، از آنجا که ماتریس جایگشت P یک ماتریس متعامد[2] است، مدل (۱) را می‌توان به صورت زیر نوشت:

(۲)

$$\min_{\mathrm{P}} ||\mathrm{A_G P} - \mathrm{P A_H}||_{\mathrm{F}}^2$$
$$\text{Subject to} \begin{cases} \sum_{\mathrm{i}} \mathrm{P_{ij}} = 1, \forall \mathrm{j} \\ \sum_{\mathrm{j}} \mathrm{P_{ij}} = 1, \ \forall \mathrm{i} \\ \mathrm{P_{ij}} \in \{0,1\}, \forall \mathrm{i,j} \end{cases}$$

۳- روش تقریبی پیشنهادی

حل دقیق مدل (۲) بازای شبکه‌های پروتئینی با تعداد پروتئین‌های زیاد یا عملاً امکان پذیر نمی‌باشد یا بسیار زمانبر است. در این بخش، مدل پیشنهادی که نسخه ساده شده و تقریبی از مدل (۲) است ارائه می‌شود. برای این منظور، از ماتریس شباهت S استفاده می‌شود که درایه سطر i–ام از ستون j–ام آن، میزان شباهت گره i–ام از گراف شبکه پروتئینی اول با گره j–ام از گراف شبکه پروتئینی دوم را مشخص می‌کند که با استفاده از روش ارائه شده در مرجع [۹]، یعنی روش ارائه شده در مقاله NETAL، بدست می‌آید. مقادیر ماتریس شباهت اعدادی میان صفر و یک هستند. مقدار نزدیک به یک، به معنای شباهت بیشتر دو پروتئین می‌باشد. لازم به ذکر است که این ماتریس، تنها طبق معیارهای توپولوژیکی ایجاد می‌شود و در محاسبه آن، از معیارهای زیستی استفاده نمی‌شود. قبلاً گفته شد که اگر در مدل (۲)، P_{ij}=1 باشد گره شماره j گراف H هم تراز با گره شماره i گراف G است که i می‌تواند هر گره از گره‌های گراف G باشد. اما براساس ماتریس شباهت می‌توان اجازه همترازی گره شماره j گراف H با فقط t تا از شبیه‌ترین گره‌های گراف G را داد. به عبارت دیگر، درصورتی که گره شماره i یکی از این شبیه‌ترین گره‌ها به گره شماره j گراف H باشد، P_{ij} می‌تواند مقدار صفر یا یک داشته باشد یعنی در این صورت، P_{ij} یکی از متغیرهای مدل محسوب می‌شود. اما درصورتی که گره شماره i یکی از این شبیه‌ترین گره‌ها به گره شماره j گراف H نباشد P_{ij} برابر با مقدار ثابت صفر خواهد بود، یعنی P_{ij} از متغیرهای مدل حذف می‌شود و مدل ساده‌تری با تعداد متغیرهای کمتری حاصل می‌شود. این مدل ساده‌تر پیشنهادی به صورت زیر می‌باشد:

(۳)

$$\min_{\mathrm{P}} ||\mathrm{A_G P} - \mathrm{P A_H}||_{\mathrm{F}}^2$$
$$\text{s.t.} \begin{cases} \sum_{\mathrm{i:S(i,j)} \geq \mathrm{t\_max(S(:,j))}} \mathrm{P_{ij}} = 1, \forall \mathrm{j} \\ \sum_{\mathrm{j:S(i,j)} \geq \mathrm{t\_max(S(\,i,:))}} \mathrm{P_{ij}} = 1, \forall \mathrm{i} \\ \mathrm{P_{ij}} \in \{0,1\}, \forall \mathrm{i,j: S(i,j)} \geq \mathrm{t\_max(S(i,:))} \\ \mathrm{P_{ij}} \in \{0,1\}, \forall \mathrm{i,j: S(i,j)} \geq \mathrm{t\_max(S(:,j))} \\ \mathrm{P_{ij}} = 0, \forall \mathrm{i,j: S(i,j)} \geq \mathrm{t\_max(S(i,:))} \\ \mathrm{P_{ij}} = 0, \forall \mathrm{i,j: S(i,j)} \geq \mathrm{t\_max(S(:,j))} \end{cases}$$

که $\mathrm{S(i,:)}$ سطر i–ام ماتریس شباهت S و $\mathrm{t\_max(S(i,:))}$، مقدار t–امین عدد بزرگتر در سطر i–ام S است. بنابراین،

[1] Frobenius Norm
[2] Orthogonal

قید پنجم، اجازه هم ترازی گره شماره i گراف G با غیر از t تا از شبیه‌ترین گره‌های گراف H را نمی‌دهد. به شکل مشابه، قید ششم، اجازه هم ترازی گره شماره j گراف H با غیر از t تا از شبیه‌ترین گره‌های گراف G را نمی‌دهد. سایر قیود مدل (۳)، ماتریس P را طوری انتخاب می‌کند که دارای ویژگی‌های یک ماتریس جایگشت باشد. البته مدل (۳)، یک مدل عدد صحیح است که حل آن هنوز هم بسیار زمانبر است. با حذف شرط صحیح بودن متغیرهای مدل، مدل ساده‌تر زیر به دست می‌آید:

$$\min_P ||A_G P - P A_H||_F^2$$
$$\text{s.t.} \begin{cases} \sum_{i:S(i,j)\geq t\_max(S(:,j))} P_{ij} = 1, \forall j \\ \sum_{j:S(i,j)\geq t\_max(S(i,:))} P_{ij} = 1, \forall i \\ P_{ij} \geq 0, \forall i,j: S(i,j) \geq t\_max(S(i,:)) \\ P_{ij} \geq 0, \forall i,j: S(i,j) \geq t\_max(S(:,j)) \\ P_{ij} = 0, \forall i,j: S(i,j) \geq t\_max(S(i,:)) \\ P_{ij} = 0, \forall i,j: S(i,j) \geq t\_max(S(:,j)) \end{cases} \quad (۴)$$

البته، جواب بهینه مدل (۴)، ضرورتاً یک ماتریس با خصوصیات ماتریس جایگشت نیست چرا که به دلیل حذف شرط صحیح بودن متغیرها، ممکن است P_{ij} مقداری غیرصحیح بین صفر یا یک باشد. برای رفع این مشکل، کافیست ماتریس P را به نحوی پیدا کنیم که تعداد عناصر غیر صفر P یا $||P||_0$ حداقل شود. چون از آنجا که جمع هر سطر و ستون این ماتریس برابر با یک است برای حداقل شدن تعداد عناصر غیر صفر P، باید فقط یک درایه از هر سطر و هر ستون برابر با یک شود و بقیه درایه‌ها برابر با صفر شوند. لذا، برای رفع مشکل غیرصحیح بودن P_{ij} ها، مدل زیر پیشنهاد می‌شود:

$$\min_P ||A_G P - P A_H||_F^2 + ||P||_0$$
$$\text{s.t.} \begin{cases} \sum_{i:S(i,j)\geq t\_max(S(:,j))} P_{ij} = 1, \forall j \\ \sum_{j:S(i,j)\geq t\_max(S(i,:))} P_{ij} = 1, \forall i \\ P_{ij} \geq 0, \forall i,j: S(i,j) \geq t\_max(S(i,:)) \\ P_{ij} \geq 0, \forall i,j: S(i,j) \geq t\_max(S(:,j)) \\ P_{ij} = 0, \forall i,j: S(i,j) \geq t\_max(S(i,:)) \\ P_{ij} = 0, \forall i,j: S(i,j) \geq t\_max(S(:,j)) \end{cases} \quad (۵)$$

البته عبارت $||P||_0$، مدل (۵) را به یک مدل نامحدب تبدیل می‌کند که حل آن دشوار است و لذا از یک روش تقریبی برای حل این مدل استفاده می‌شود. برای این منظور، از ایده مرجع [۱۴] استفاده می‌شود. براساس این ایده، برای حل مدل $\min_{P\in\wp}||P||_0$، نخست مسئله زیر حل می‌شود:

(۶)

$$\min_{P\in\wp} \sum_{ij} \lambda_{ij} P_{ij}$$

که مقدار λ_{ij} برابر با یک در نظر گرفته می‌شود. پس از حل مدل (۶) و به‌دست آوردن مقدار بهینه P در این مدل، مقدار λ_{ij} طبق رابطه زیر به‌روز رسانی می‌گردد و این روند تا همگرایی تکرار می‌شود:

(۷)

$$\lambda_{ij} = \begin{cases} \infty & P_{ij} \leq \varepsilon, \\ \dfrac{1}{P_{ij}} & \text{othertwise.} \end{cases}$$

بنابراین، شبه کد الگوریتم پیشنهادی برای حل مدل (۵) بصورت زیر می‌باشد:

۱. قرار بده $\lambda_{ij} = 1$

۲. مدل زیر را که یک مدل درجه دوم استاندارد است با الگوریتم active set [۱۵] حل کن که در ضمیمه توضیح داده شده است:

$$\min_P ||A_G P - P A_H||_F^2 + \sum_{ij} \lambda_{ij} P_{ij}$$
$$\text{s.t.} \begin{cases} \sum_{i:S(i,j)\geq t\_max(S(:,j))} P_{ij} = 1, \forall j \\ \sum_{j:S(i,j)\geq t\_max(S(i,:))} P_{ij} = 1, \forall i \\ P_{ij} \geq 0, \forall i,j: S(i,j) \geq t\_max(S(i,:)) \\ P_{ij} \geq 0, \forall i,j: S(i,j) \geq t\_max(S(:,j)) \\ P_{ij} = 0, \forall i,j: S(i,j) \geq t\_max(S(i,:)) \\ P_{ij} = 0, \forall i,j: S(i,j) \geq t\_max(S(:,j)) \end{cases}$$

۳. مقدار λ_{ij} را براساس رابطه (۷) بروز کن.

۴. اگر λ_{ij}های این مرحله، نسبت به λ_{ij} مرحله قبل تغییر کرده‌اند برو به مرحله ۲.

٤- آزمایش‌ها

در این بخش، ابتدا معیارهای ارزیابی روشهای هم‌ترازی معرفی می‌شود. سپس تعدادی شبکه پروتئینی مصنوعی و واقعی معرفی می‌شوند. در نهایت، براساس این شبکه‌های پروتئینی، روش پیشنهادی ارزیابی شده و با مقاله NETAL نیز مقایسه می‌شود.

لازم به ذکر است که تمامی آزمایشات روی سیستم ٦٤ بیتی، با پردازنده Intel core i5 و هشت گیگا بایت حافظه انجام شده است و هر روش هم ترازی در محیط MATLAB R2013a پیاده سازی گردیده است.

٤-١- معیارهای ارزیابی کیفیت هم‌ترازی

• *درستی یال*[1]*:*

یک معیار استاندارد برای اندازه گیری کیفیت توپولوژیکی، درستی یال یا EC است و طبق تعریف، برابر است با نسبت یال‌هایی از شبکه اول G_1 که با یال‌های شبکه دوم G_2 هم تراز شده‌اند [١٦]. درستی یال به‌صورت زیر محاسبه می‌شود:

$$EC = \frac{|\{(u,v)\epsilon E_1 : (f(u),f(v))\epsilon E_2|}{|E_1|}$$

که f تابع یک به یک همترازی بدست آمده با استفاده از یک روش هم ترازی است. لذا، صورت کسر رابطه (١٠)، تعداد یال‌های هم‌تراز شده و مخرج کسر، تعداد کل یال های شبکه اول را نشان می‌دهد. درستی یال، عددی بین صفر و یک است و هرچه درستی یال به یک نزدیک‌تر باشد، دو شبکه از لحاظ توپولوژیکی به یکدیگر شبیه‌ترند. در صورتی که درستی یال برابر یک باشد، یعنی اگر EC=1 باشد، هم‌ترازی انجام شده توانسته است شبکه اول را به صورت زیرگراف یک‌ریخت[2] در شبکه دوم نگاشت کند که این بیشترین شباهت توپولوژیکی ممکن می‌باشد. این معیار را به‌صورت درصد نیز بیان می‌نمایند.

• *فاصله هم‌ترازی:*

به منظور ارزیابی مدل مطرح شده، علاوه بر معیارهای درستی یال، معیار فاصله هم‌ترازی نیز مورد استفاده قرار می‌گیرد که بصورت زیر محاسبه می‌شود:

$$D(P) = ||A_G P - P A_H||_F^2$$

هرچه فاصله همترازی کوچک‌تر باشد، هم‌ترازی بهتری صورت گرفته است.

٤-٢- شبکه‌های پروتئینی مورد آزمایش

• *شبکه‌های پروتئینی واقعی:*

در این مقاله، از شبکه‌های واقعی گونه‌های کرم و موش پایگاه داده Isobase و همچنین شبکه‌های مخمر و باکتری پایگاه داده MINT [١٧] استفاده شده است. مشخصات شبکه‌های مذکور مطابق جدول ١ می‌باشد.

جدول ١- اطلاعات شبکه‌های پروتئینی کرم، موش، مخمر و باکتری

گونه	مخفف	تعداد پروتئین	تعداد برهمکنش
موش (C.elegans)	Ce	٢٨٠٥	٤٤٩٥
کرم (M.musculus)	Mm	٢٩٠	٢٤٢
مخمر (Saccharomyces-cerevisiae)	Sc	٢٣٩٠	١٦١٢٧
باکتری (Escherichia-coli)	Ec	٧٠١	١٦٤٠

• شبکه‌های پروتئینی مصنوعی

از آن‌جا که در اغلب مقالات مرتبط، نسبت به هم ترازی شبکه پروتئینی انسان (H.Sapiens:Hs) با گونه‌های دیگر، مخصوصاً مخمر اقدام کرده‌اند، در این تحقیق نیز با استفاده از روش پیشنهادی نسبت به هم‌ترازی شبکه پروتئینی مخمر و انسان نیز اقدام می‌شود. اما به دلیل اینکه تعداد پروتئین‌های شبکه انسان بسیار زیاد و حدود ٩١٤١ پروتئین می‌باشد و امکانات سخت افزاری موجود برای اجرای این تعداد پروتئین در شبکه در دسترس نبود، به جای استفاده

[1] Edge Correctness(EC)
[2] Isomorphic

از شبکه واقعی انسان، از شبکه مصنوعی انسان و مخمر مخزن NAPAbench [۱۸] استفاده می‌شود که تعداد پروتئین کمتری نسبت به شبکه‌های واقعی دارند. لازم به ذکر است که این شبکه‌های مصنوعی برخلاف شبکه‌های پروتئینی واقعی، فاقد نویز هستند. مشخصات این دو شبکه پروتئینی مصنوعی مطابق جدول ۲ می‌باشد.

جدول ۲- مشخصات شبکه‌های پروتئینی مصنوعی انسان و مخمر

شبکه پروتئینی مصنوعی	تعداد پروتئین (گره)			
انسان (Hs)	۱۵۰۰	۳۰۰۰	۳۵۰۰	٤۰۰۰
مخمر (Cs)	۹۰۰	۲۰۰۰	۲۵۰۰	۳۰۰۰

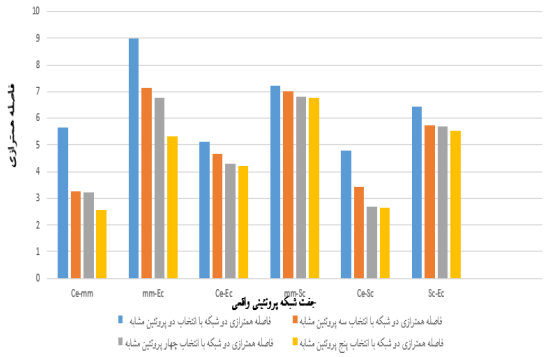
٤-۳- نتایج آزمایش‌ها

٤-۳-۱- ارزیابی روش پیشنهادی

یکی از عوامل تاثیرگذار در معیار درصد درستی یال، معیار فاصله هم ترازی و معیار مدت زمان اجرای هم ترازی در روش پیشنهادی، پارامتر t است. در روش پیشنهادی، فقط به یکی از t گره از شبیه‌ترین گره‌های شبکه پروتئینی اول، اجازه هم تراز شدن با یک گره از شبکه پروتئینی دوم داده می‌شود که می‌تواند بر روی دقت هم‌ترازی که با معیارهای درصد درستی یال و فاصله هم‌ترازی سنجیده می‌شود، تأثیر منفی بگذارد. هر چه مقدار t کوچک‌تر باشد، تعداد متغیرهای مدل پیشنهادی کمتر می‌شود و مدت زمان لازم برای اجرای مدل کاهش می‌یابد. اگر مقدار t برابر با حداکثر تعداد گره‌ها باشد، مدل پیشنهادی، به مدل هم ترازی دقیق (۲) تبدیل می‌شود که زمان اجرای آن به دلیل تعداد متغیرهای بیشتر، افزایش می‌یابد و البته دقت هم ترازی نیز افزایش می‌یابد.

• *ارزیابی روش پیشنهادی براساس معیار فاصله هم‌ترازی بر روی شبکه‌های واقعی:*

شکل ۱ فاصله هم‌ترازی میان ٦ جفت شبکه پروتئینی واقعی بدست آمده از ترکیب هر جفت شبکه پروتئینی موجود در جدول ۱ را نشان می‌دهد زمانی که فقط به دو، سه، چهار یا پنج پروتئین شبیه‌تر اجازه هم ترازی داده می‌شود (t=2,3,4,5). همان‌گونه که در این شکل مشاهده می‌شود زمانی که به تعداد بیشتری گره شبیه‌تر اجازه هم ترازی داده می‌شود، فاصله هم‌ترازی شبکه‌ها کاهش می‌یابد.

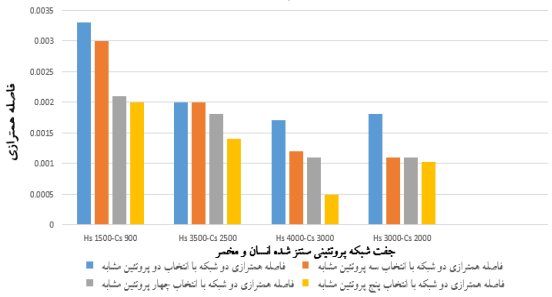


شکل ۱- فاصله هم‌ترازی شش جفت شبکه پروتئینی واقعی (هر دو جفت از شبکه‌های پروتئینی جدول ۱) زمانی که فقط به دو، سه، چهار و پنج پروتئین شبیه‌تر اجازه هم ترازی داده می‌شود (t=2,3,4,5).

• *ارزیابی روش پیشنهادی براساس معیار فاصله هم‌ترازی بر روی شبکه‌های مصنوعی:*

شکل ۲ فاصله هم‌ترازی میان شبکه‌های پروتئینی انسان و مخمر در اندازه‌های مختلف ذکر شده در جدول ۲ را نشان می‌دهد زمانی که فقط به دو، سه، چهار یا پنج پروتئین شبیه‌تر اجازه هم‌ترازی داده می‌شود (t=2,3,4,5). همان‌گونه که در این نمودار مشاهده می‌شود زمانی که به تعداد بیشتری گره شبیه‌تر اجازه هم‌ترازی داده می‌شود، فاصله هم‌ترازی شبکه‌ها کاهش می‌یابد.

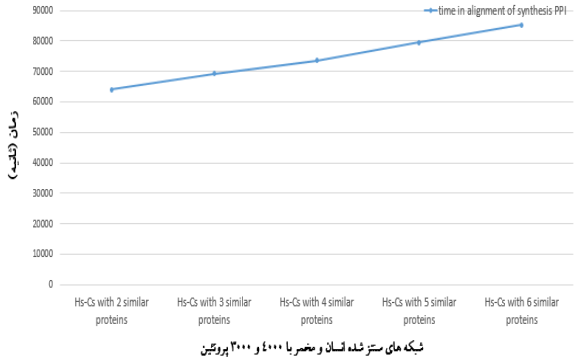
از مقایسه شکل ۱ و ۲ مشاهده می‌شود که فاصله هم‌ترازی به‌دست آمده با روش پیشنهادی بر روی شبکه‌های مصنوعی نسبت به شبکه‌های واقعی بسیار کم‌تر است و چیزی نزدیک به صفر شده است که به دلیل عدم وجود نویز در شبکه‌های پروتئینی مصنوعی است.



شکل ۲- فاصله هم‌ترازی میان شبکه‌های پروتئینی انسان و مخمر در اندازه‌های مختلف ذکر شده در جدول ۲ را نشان می‌دهد زمانی که فقط به دو، سه، چهار و پنج پروتئین شبیه‌تر اجازه هم ترازی داده می‌شود (t=2,3,4,5).

• *ارزیابی روش پیشنهادی براساس مدت زمان اجرای هم‌ترازی بر روی شبکه‌های پروتئینی مصنوعی:*

شکل ۳، مدت زمان اجرای هم‌ترازی میان بزرگترین شبکه پروتئینی انسان و مخمر ذکر شده در جدول ۲ با استفاده از روش پیشنهادی را نشان می‌دهد زمانی که فقط به دو، سه، چهار، پنج یا شش پروتئین شبیه‌تر اجازه هم ترازی داده می‌شود (t=2,3,4,5,6). همان‌گونه که در این نمودار مشاهده می‌شود زمانی که به تعداد بیشتری گره شبیه‌تر اجازه هم ترازی داده می‌شود، مدت زمان اجرای هم‌ترازی شبکه‌ها افزایش می‌یابد چرا که تعداد متغیرهای مدل پیشنهادی، افزایش می‌یابد.

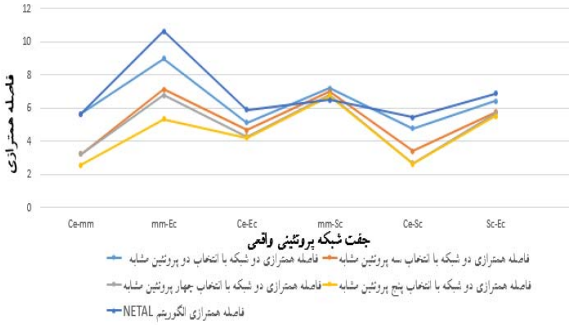


شکل ۳- مدت زمان اجرای هم‌ترازی بزرگترین شبکه مصنوعی انسان و مخمر ذکرشده در جدول ۲ با استفاده از روش پیشنهادی، زمانی که فقط به دو، سه، چهار، پنج یا شش پروتئین شبیه‌تر (t=2,3,4,5,6) اجازه هم ترازی داده می‌شود..

۲-٤-۳- مقایسه روش پیشنهادی

• *مقایسه روش پیشنهادی با الگوریتم NETAL[۹] براساس معیار فاصله هم‌ترازی:*

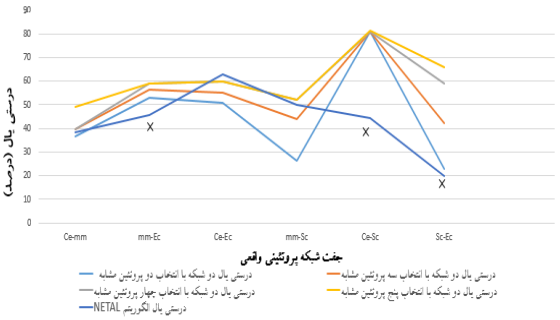
شکل ٤ روش پیشنهادی و الگوریتم NETAL را براساس معیار فاصله هم‌ترازی و به ازای شش جفت شبکه پروتئینی واقعی مقایسه می‌کند. همان‌طور که ملاحظه می‌شود در ۵ مورد از شش مورد هم‌ترازی، فاصله هم‌ترازی روش پیشنهادی زمانی که فقط به دو، سه، چهار یا پنج پروتئین شبیه‌تر اجازه هم ترازی داده می‌شود (t=2,3,4,5)، کمتر از روش NETAL می‌باشد.



شکل ٤- مقایسه روش پیشنهادی با الگوریتم NETAL[۹] براساس معیار فاصله همترازی.

• *مقایسه روش پیشنهادی با الگوریتم NETAL[۹] براساس معیار درستی یال:*

شکل ۵ روش پیشنهادی و الگوریتم NETAL را براساس معیار درصد درستی یال و به ازای شش جفت شبکه پروتئینی واقعی مقایسه می‌کند. همان‌طور که ملاحظه می‌شود، در ۵ مورد از شش مورد همترازی، درصد درستی یال روش پیشنهادی زمانی که فقط به پنج پروتئین شبیه‌تر اجازه هم ترازی داده می‌شود (t=5)، بیشتر از روش NETAL می‌باشد.



شکل ۵–مقایسه روش پیشنهادی با الگوریتم NETAL[۹] براساس معیار درصد درستی یال.

۵- بحث و نتیجه‌گیری

چالش‌های اصلی مسئله هم‌ترازی شبکه‌های پروتئینی، زمان‌بر بودن حل مساله به دلیل وجود پروتئین‌های بسیار زیاد در شبکه‌های پروتئینی واقعی و همچنین NP-complete بودن این مساله است. در این مقاله، برای کاهش پیچیدگی حل مدل ریاضی مساله هم ترازی، تقریبی از مدل ارائه شد. به عبارت دقیق‌تر، از ماتریس شباهت دو شبکه پروتئینی برای کاهش تعداد متغیرهای مدل استفاده گردید. برای این منظور، در مدل تقریبی پیشنهادی، به‌جای بررسی هم‌ترازی هر یک از پروتئین‌های شبکه نخست با تمام پروتئین‌های شبکه دوم، هم‌ترازی هر یک از پروتئین‌های شبکه نخست فقط با شبیه‌ترین پروتئین‌های شبکه دوم بررسی می‌شود. در نهایت، مدل پیشنهادی، برای هم‌ترازی شبکه‌های پروتئینی واقعی گونه‌های مختلف و نیز شبکه‌های مصنوعی آزمایش شد. نتایج تجربی، نشان دهنده بهبود دقت هم‌ترازی شبکه‌ها نسبت به روش تقریبی NETAL و نیز کاهش زمان اجرا، نسبت به مدل دقیق می‌باشد. در ضمن، روش پیشنهادی، توانسته است بر روی شبکه‌های مصنوعی فاقد نویز، به دقت بسیار مطلوب دست یابد.

ضمیمه- الگوریتم active-set [۱۵]

مساله درجه ۲ با قیود خطی زیر را درنظر بگیرید:

(۸)

$$\min_x \frac{1}{2}x^T Qx - c^T x$$
$$\text{Subject to } Ax \geq b.$$

که $x \in R^n$ و Q ماتریس متقارن و A ماتریس $m \times n$ و b بردار m مؤلفه‌ای است.

فرض کنید که $x^{(k)}$ یک جواب ممکن[1] مساله در تکرار k–ام باشد که در قیود مساله صدق می‌کند (که با حل یک مساله خطی، یعنی همان مساله (8) بدون تابع هدف، بدست می‌آید).

مجموعه کاری[2] یعنی مجموعه قیود فعال[3] در تکرار جاری را بصورت زیر تعریف می‌کنیم:

$$W = \{i | a_i^T x^{(k)} = b_i\}$$

که a_i، ستون i–ام ماتریس معلوم A و b_i، مؤلفه i–ام بردار b است.

فرض کنید که قیود فعال بازای $x^{(k)}$ همان قیود فعال به ازای جواب بهینه x^* باشند. بنابراین به جای حل مساله (8)، کافیست مساله زیر حل شود:

(۹)

$$\min_p \frac{1}{2}(x^{(k)} + p)^T Q(x^{(k)} + p) - c^T(x^{(k)} + p)$$
$$\text{Subject to} A_W p = 0.$$

که $x^{(k)} + p$ جواب جدید مساله درنظرگرفته شده است و A_W سطرهایی از A است که توسط مجموعه W مشخص شد. تابع لاگرانژ مساله (9) به صورت زیر است:

$$L = \frac{1}{2}(x^{(k)} + p)^T Q(x^{(k)} + p) - c^T(x^{(k)} + p) - \lambda^T(A_W p).$$

که λ ضرایب لاگرانژ مساله است. در نقطه بهینه دوگان مساله (9) داریم:

(۱۰)

$$0 = \frac{\partial L}{\partial p} = Qp - \lambda^T A_W + Qx^{(k)} - c.$$

$$0 = \frac{\partial L}{\partial \lambda} = A_W p.$$

[1] Feasible
[2] Working set
[3] Active constraint

بنابراین، پس از حل دستگاه معادلات خطی فوق، قرار می دهیم:

$$x = x^{(k)} + p.$$

اگر $Ax \geq b$ آنگاه x، جواب بهینه مساله (8) نیز هست و پایان الگوریتم، اعلام می‌شود. درغیراین صورت، بزرگ‌ترین گام α را به نحوی تعیین می‌کنیم که

$$A_W(x^{(k)} + \alpha p) \geq b,$$

یعنی قرار می‌دهیم

$$\alpha = \min_{i|a_i^T p<0} \frac{a_i^T x^{(k)} - b_i}{-a_i^T p}$$

سپس قرار می‌دهیم

$$x^{(k+1)} = x^{(k)} + \alpha p,$$

$$W = W \cup \{I\},$$

که $a_I^T x^{(k+1)} = b_I$. یعنی، قید شماره i به مجموعه قیود فعال، اضافه می‌شود.

درضمن، اگر پس از حل دستگاه معادلات ()، $\lambda > 0$ شود آنگاه جواب بدست آمده، جواب بهینه مساله زیر نیز هست:

(۱۱)

$$\min_p \frac{1}{2}(x^{(k)} + p)^T Q(x^{(k)} + p) - c^T(x^{(k)} + p)$$
$$\text{Subject to} A_W(x^{(k)} + p) \geq b_W.$$

لذا قرار می‌دهیم

$$x^{(k+1)} = x^{(k)} \alpha p.$$

ولی اگر به ازای برخی از مقادیر $\lambda_i < 0, i$ شود آنگاه

$$a_i^T(x^{(k)} + p) < b_i.$$

لذا، قرار می‌دهیم

$$W = W - \{i\}.$$
$$x^{(k+1)} = x^{(k)} + p.$$

یعنی، قید شماره i از مجموعه قیود فعال، حذف می‌شود. این الگوریتم تا همگرایی، ادامه می‌یابد.

٦- مراجع

[1]R. Sharan and T. Ideker, "Modeling cellular machinery through biological network comparison," *Nature biotechnology,* vol. 24, pp. 427-433, 2006.

[2]V. Saraph and T. Milenković, "MAGNA: maximizing accuracy in global network alignment," *Bioinformatics,* vol. 30, pp. 2931-2940, 2014.

[3]A. Zhang, *Protein interaction networks: computational analysis*: Cambridge University Press, 2009.

[4]M. El-Kebir, J. Heringa, and G. W. Klau, "Natalie 2.0: Sparse Global Network Alignment as a Special Case of Quadratic Assignment," *Algorithms,* vol. 8, pp. 1035-1051, 2015.

[5]H. Almohamad and S. O. Duffuaa, "A linear programming approach for the weighted graph matching problem," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 15, pp. 522-525, 1993.

[6]A. E. Aladağ and C. Erten, "SPINAL: scalable protein interaction network alignment," *Bioinformatics,* vol. 29, pp. 917-924, 2013.

[7]J. Crawford, Y. Sun, and T. Milenković, "Fair evaluation of global network aligners," *Algorithms for Molecular Biology,* vol. 10, p. 1, 2015.

[8]M. Gong, Z. Peng, L. Ma, and J. Huang, "Global Biological Network Alignment by Using Efficient Memetic Algorithm," 2015.

[9]B. Neyshabur, A. Khadem, S. Hashemifar, and S. S. Arab, "NETAL: a new graph-based method for global alignment of protein–protein interaction networks," *Bioinformatics,* vol. 29, pp. 1654-1662, 2013.

[10]O. Kuchaiev, T. Milenković, V. Memišević, W. Hayes, and N. Pržulj, "Topological network alignment uncovers biological function and phylogeny," *Journal of the Royal Society Interface,* vol. 7, pp. 1341-1354, 2010.

[11]O. Kuchaiev and N. Pržulj, "Integrative network alignment reveals large regions of global network similarity in yeast and human," *Bioinformatics,* vol. 27, pp. 1390-1396, 2011.
[12]R. Singh, J. Xu, and B. Berger, "Pairwise global alignment of protein interaction networks by matching neighborhood topology," in *Annual International Conference on Research in Computational Molecular Biology*, 2007, pp. 16-31.
[13]M. Zaslavskiy, F. Bach, and J.-P. Vert, "A path following algorithm for the graph matching problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 31, pp. 2227-2242, 2009.
[14]K. Huang, D. Zheng, J. Sun, Y. Hotta, K. Fujimoto, and S. Naoi, "Sparse learning for support vector classification," *Pattern Recognition Letters,* vol. 31, pp. 1944-1951, 2010.
[15]E. Wadbro. (016). *Quadratic programs and Active set methods*. Available: http://people.cs.umu.se/eddiew/optpde2016/QP.pdf
[16]R. Patro and C. Kingsford, "Global network alignment using multiscale spectral signatures," *Bioinformatics,* vol. 28, pp. 3105-3114, 2012.
[17]T. Bruckdorfer, O. Marder, and F. Albericio, "From production of peptides in milligram amounts for research to multi-tons quantities for drugs of the future," *Current pharmaceutical biotechnology,* vol. 5, pp. 29-43, 2004.
[18]S. M. E. Sahraeian and B.-J. Yoon, "A network synthesis model for generating protein interaction network families," *PloS one,* vol. 7, p. e41474, 2012.