DOR: 20.1001.1.23832738.1401.35.1.10.4

مقاله پژوهشی

تأثیر الگوریتم‌های فراابتکاری در همترازی شبکه‌های مبتنی بر میانکنش پروتئین-پروتئین در پنج گونه زیستی

الهام مهدی‌پور و محمد قاسم‌زاده*

ایران، یزد، دانشگاه یزد، دانشکده مهندسی کامپیوتر

تاریخ دریافت: ۱۳۹۹/۰۵/۲۵ تاریخ پذیرش: ۱۳۹۹/۰۷/۰۷

چکیده

از طریق همترازی توالی ژنوم می‌توان دانش زیستی گونه‌های مختلف را به نواحی حفاظت شده‌ی توالی انتقال داد. به‌طور مشابه، از طریق همترازی شبکه زیستی، می‌توان دانش نواحی حفاظت شده‌ی شبکه‌های مولکولی را به نواحی مختلف حفاظت شده‌ی گونه‌های متفاوت انتقال داد. لذا با تکیه بر همترازی شبکه‌های زیستی می‌توان «همسانی مبتنی بر توالی» را به «همسانی مبتنی بر شبکه» تعمیم داد. کشف همترازی شبکه‌ها به جهت کاربردهای آن، مانند کشف داروهای جدید، ردیابی روند پیشرفت بیماری‌ها و یا پیش‌بینی رفتار کاربران در شبکه‌های اجتماعی، از اهمیت ویژه‌ای برخوردار است. در این رابطه، چالش اصلی این است که یافتن همترازی‌های موجود در دو شبکه، یک مسئله‌ی از مرتبه‌ی «اِن پی-سخت» است. در چنین وضعیتی از راه‌حل‌های تقریبی مانند الگوریتم‌های فراابتکاری که نسبتاً سریع هستند، بهره می‌گیریم. بخش اصلی این پژوهش، مقایسه الگوریتم‌های همترازی شبکه از دیدگاه معیارهای ارزیابی مربوطه، زمان اجرا، میزان مصرف حافظه و میزان پیچیدگی شبکه‌های مورد تست می‌باشد. نتایج آزمایشی از اجرای جدیدترین و مشهورترین الگوریتم‌های مرتبط بر روی مجموعه داده‌ی شبکه‌های زیستی بیوگرید به‌دست‌آمده‌اند. نتایج پیاده‌سازی و ارزیابی حاکی از آن است که با بهره‌گیری از الگوریتم‌های فراابتکاری ژنتیک، میمتیک، بهینه-سازی توده ذرات، تبرید شبیه‌سازی شده و کلونی مورچگان می‌توان به نتایج ارزشمندی دست یافت. روش‌های یادشده با بکارگیری توابع مکاشفه‌ای مناسب، تنها بخش‌های کوچکی از داده‌های قابل جستجو را مورد بررسی قرار می‌دهند، لذا غالباً موفق به کشف پاسخ بهینه و یا قابل‌قبول در زمان کوتاهی می‌شوند.

واژه‌های کلیدی: الگوریتم‌های فراابتکاری، تعامل پروتئین-پروتئین، تطبیق گراف، زیرگراف ایزومورف، همترازی شبکه.

* نویسنده مسئول، تلفن: ۳۱۲۳۲۳۵۹-۰۳۵ ، پست الکترونیکی: m.ghasemzadeh@yazd.ac.ir

مقدمه

یکی از مشکلات جدی در مطالعات زیستی، بکارگیری داده‌های زیستی در حجم‌های انبوه است (۳). الگوریتم‌های جستجو یکی از شاخه‌های مهم در تحقیقات و بهینه‌سازی هستند که بعلت بزرگتر و پیچیده‌تر شدن مسائل، نیاز به روش‌های جستجوی کارآمدتر، بیشتر احساس می‌شود. الگوریتم‌های فراابتکاری الهام‌گرفته از طبیعت هستند که در حل مسائل بهینه‌سازی دشوار بسیار کارآمد عمل می‌کنند.

هنگامی که اندازه مسئله بزرگ باشد، استفاده از روش‌های جستجوی ناآگاهانه میسر نخواهد بود. چرا که در این موارد، دامنه مسئله چنان بزرگ می‌باشد که رسیدن به جواب مسئله در مدت زمان قابل قبول با استفاده از این روش‌های جستجو غیر ممکن خواهد بود. از این رو نیاز به روش‌هایی داریم که بتواند در بخش‌های مختلف فضای داده جستجو را انجام دهد. جستجوی فراابتکاری نام خانواده‌ای از الگوریتم‌های جستجوی آگاهانه است که از یک پدیده‌ی طبیعی، بعنوان مدل برای بهینه سازی کاوش

DOR: 20.1001.1.23832738.1401.35.1.10.4
مقاله پژوهشی

هوشمندانه فضای جستجوی داده‌های بسیار بزرگ در مسائل پیچیده استفاده می‌شود.

همترازی شبکه‌های زیستی (Biological network alignment) یکی از مسائل دشواری می‌باشد که توجه محققین کامپیوتر، زیست‌شناسی و بیوانفورماتیک را به خود جلب کرده است. شبکه زیستی به شبکه‌ای گفته می‌شود که در مورد سیستم‌های زیستی مورد استفاده قرار می‌گیرد. یکی از این شبکه‌های زیستی، شبکه‌های مبتنی بر میانکنش پروتئین-پروتئین (PPI= Protein-Protein Interaction) هستند؛ که از پروتئین‌ها و ارتباط بین آن‌ها تشکیل شدند.

از دیدگاه شیمی، پروتئین ترتیب خطی از اسیدآمینه‌ها است که زنجیره پلی‌پپتید هم نامیده می‌شود (۲۲). پروتئین‌ها می‌توانند باتوجه به وجود ساختارهای مکمل و ویژگی‌های فیزیکی اسیدآمینه‌ها به یکدیگر متصل شده و ساختارهای دسته‌جمعی پیچیده‌ای از پروتئین‌ها را تشکیل دهند. تعاملات بسیار پروتئین‌ها داخل یک سلول با یکدیگر شبکه تعاملی پروتئین‌ها را تشکیل می‌دهد که در این شبکه گره‌ها بیانگر پروتئین‌ها و یال‌ها بیانگر تعاملی یا ارتباطی است که پروتئین‌ها برقرار می‌کنند.

تحقیقات زیادی بر روی تعیین مشترکات و انتقال حاشیه-نویسی بین شبکه‌های PPI گونه‌های مختلف متمرکز شده-اند که اغلب توسط همترازی شبکه انجام می‌شود (۲۲). هدف از همترازی شبکه، شناسایی نگاشت (Mapping) بین پروتئین‌ها در شبکه PPI است که می‌توانند بیانگر شباهت‌های توپولوژی، عملکردی، و نواحی حفاظت شده تکاملی در شبکه‌های PPI باشند. الگوریتمی که این نگاشت را شناسایی می‌کند، همترازکننده (Aligner) نامیده می‌شود. همترازی شبکه‌ها باعث کشف اطلاعات ارزشمندی همچون مسیرهای حفاظت شده تکاملی و پیچیدگی‌های پروتئین شده است.

عملکرد بسیاری از پروتئین‌ها بصورت کامل شناسایی نشده است (۲۲)، که بدلیل این مسئله چالش‌های شناسایی نقش آنها در فرآیندهای زیستی و بیماری‌ها مشکلاتی ایجاد می‌نماید. بنابراین، این فرآیندها و بیماری‌ها بصورت آزمایشی در ارگانیسم‌های مدل، مانند *مخمر* و *موش* مورد بررسی قرار می‌گیرند و دانش مربوطه از موجودات مدل با استفاده از شباهت‌های شبکه به *انسان* تعمیم داده می‌شود (۲۲). لذا همترازی با کمک به انتقال اطلاعات عملکردی بین گونه‌های مختلف، به نوبه خود برای پیش‌بینی مکانیزم‌های بیماری *انسان* و یا پیش‌بینی فرآیند پیری *انسان* مورد استفاده قرار می‌گیرد(۱۱).

در این پژوهش، کارآمدی الگوریتم‌های فراابتکاری در حل سریع و بهینه مسئله همترازی بر شبکه‌های مبتنی بر میانکنش پروتئین-پروتئین مورد بررسی قرار داده می‌شود. همچنین در این پژوهش به سوالات زیر پاسخ داده می‌شود: ۱) کدام روش همترازی شبکه را دانشمندان زیست‌شناسی استفاده می‌کنند؟ ۲) چگونه می‌توان دو همترازی متفاوت را مقایسه نمود؟ ۳) چگونه کیفیت همترازی را اندازه‌گیری نماییم؟

همترازی شبکه: همترازی شبکه به معنای یافتن بهترین راه برای متناظر نمودن یک شبکه درون شبکه دیگر است. این امر در چندین حوزه شامل تطبیق هستی‌شناسی (Ontology)، شناسایی الگو، پردازش زبان و شبکه‌های اجتماعی کاربرد دارد (۲۵). بنابراین هدف همترازی شبکه به میزان زیادی به محتوای شبکه بستگی دارد. روش‌های گوناگونی جهت همترازی شبکه وجود دارد. این تنوع به واسطه پیچیدگی محاسبات همترازی شبکه است (۱۳).

تعاریف اولیه: شبکه یا گراف، زوج مرتب $N=(V,E)$ است که V مجموعه گره‌ها و E مجموعه یال‌ها است که برخی گره‌های V را به یکدیگر متصل می‌کند. تعداد شبکه-های ممکن با n گره برابر 2^{n^2} می‌باشد، این مقدار نمایی باعث می‌شود دسته‌بندی شبکه و مسئله تطبیق آن از نظر محاسباتی دشوار باشند.

DOR: 20.1001.1.23832738.1401.35.1.10.4
مقاله پژوهشی

مسئله تطبیق شبکه به تئوری گراف و مسئله زیرگراف ایزومورف تعمیم می‌یابد، و این پرسش را مطرح می‌نماید که آیا شبکه $N_1=(V_1,E_1)$ زیرگراف دقیقی از شبکه $N_2=(V_2,E_2)$ هست یا خیر. ثابت شده است که این مسئله جزء «ان پی-کامل» محسوب می شود به این معنی که راه‌حل کارآمد و چندجمله‌ای ندارد. همچنین ثابت شده است که مسئله تطبیق شبکه مسئله «ان پی-سخت» است (۲۲،۱۶). بنابراین همترازی شبکه، کلی‌تر از یافتن مناسب-ترین زیرشبکه N_1 درون N_2 است حتی اگر N_1 زیرگرافی دقیقی از N_2 نباشد (۲۲).

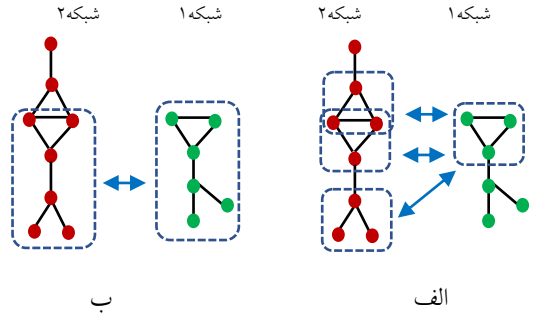
همترازی شبکه دربرگیرنده دو بحث اصلی است: همترازی موضعی (Local alignment) و همترازی سراسری (Global alignment).

هدف اصلی همترازی موضعی این است که نواحی کوچک دقیقاً همتراز شوند (۱۳). در مقابل هدف همترازی شبکه سراسری تولید نگاشت یک به یک بین گره‌های دو شبکه است. اکثر تحقیقات اخیر بر روی همترازی شبکه سراسری متمرکز شده‌اند (۱۳). هدف همترازی سراسری این است که نگاشتی از پروتئین‌های شبکه کوچکتر به پروتئین‌های شبکه بزرگتر پیدا شود. بطور مرسوم، بین دو شبکه PPI، $N_1=(V_1,E_1)$ و $N_2=(V_2,E_2)$ با فرض $|V_1|\le|V_2|$، همترازی f نگاشت بین گره‌های V_1 و گره‌های V_2 است:

$$f:V_1^{'}\to V_2^{'} \qquad (۱)$$

بطوریکه $V_1^{'}\subseteq V_1$ و $V_2^{'}\to V_2$.

در همترازی موضعی، $V_1^{'}$ و $V_2^{'}$ زیرمجموعه‌های کوچک متناظر با مشابه‌ترین گره‌ها در شبکه‌ها هستند (شکل۱، قسمت الف)؛ در حالی که هدف همترازی سراسری، همترازی همه گره‌های V_1 است (شکل ۱، قسمت ب). از این رو، همترازی موضعی نگاشت‌های چند-به-چند تولید می‌کند یعنی یک گره از $V_1^{'}$ می‌تواند به چندین گره در $V_2^{'}$ نگاشت شود و برعکس؛ همترازی سراسری نگاشت های یک-به-یک تولید می‌کند یعنی همترازی دوطرفه بین $V_1^{'}$ و $V_2^{'}$ (۲۲). شکل۱ تفاوت بین همترازی موضعی و سراسری را نشان می‌دهد. همان‌گونه که در شکل۱ قسمت الف مشاهده می‌شود در همترازی موضعی، یک بخش از شبکه اول می‌تواند به چندین بخش از شبکه دوم نگاشت شود؛ اما در همترازی سراسری همانند شکل ۱ قسمت ب، کل یا بخشی از شبکه اول، فقط به یک بخش از شبکه دوم نگاشت می‌شود.



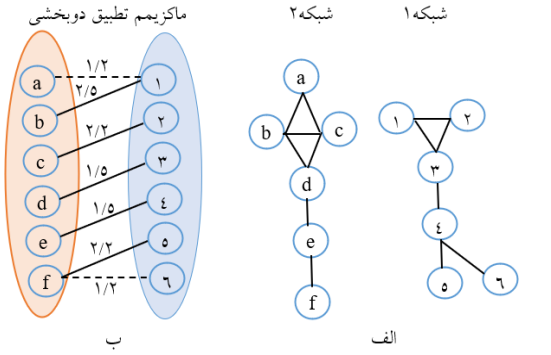
شکل ۱- همترازی شبکه، الف: شبکه همترازی موضعی ، ب: شبکه همترازی سراسری

بدلیل «ان پی-کامل» بودن مسئله زیرگراف ایزومورف، همترازکننده‌های شبکه از توابع مکاشفه‌ای (راه‌حل‌های تقریبی) استفاده می‌کنند که برای تقریب راه‌حل مسئله همترازی شبکه، تابع هدف و الگوریتم‌های بهینه‌سازی متفاوتی دارند. برای کشف همترازی‌هایی که تابع هدفشان ماکزیمم‌سازی است، اغلب همترازکننده‌ها یا به گراف‌های همترازی اعتماد می‌کنند یا از تطبیق دوبخشی استفاده می کنند (۲۲).

- **گراف همترازی**: بطور کلی، جستجو برای همترازی با امتیاز بالا متناظر با جستجو در گراف همترازی برای یافتن زیرگرافی با بزرگترین مجموع گره و وزن یال‌ها است.
- **گراف دوبخشی**: گراف $B=(V_1\bigcup V_2,E)$ دوبخشی است اگر مجموعه گره‌های آن به دو مجموعه V_1 و V_2 تقسیم شود، بطوریکه یال‌ها (در E) فقط از گره‌های V_1 به

DOR: 20.1001.1.23832738.1401.35.1.10.4
مقاله پژوهشی

گره‌های V_2 متصل باشند. وزن‌ها می‌توانند متناظر با هزینه-های نگاشت گره به یال‌ها نسبت داده شوند. بسیاری از همترازکننده‌های سراسری رویکردی دو مرحله‌ای را استفاده می‌کنند: آن‌ها ابتدا امتیاز نگاشت بین گره‌های شبکه‌های همترازشده را محاسبه نموده (شکل۲-الف دو شبکه کاندید برای همترازی را نشان می‌دهد)، سپس مدل همترازی‌شان را همانند وزن تطبیق دوبخشی ماکزیمم می‌نمایند (شکل۲-ب).



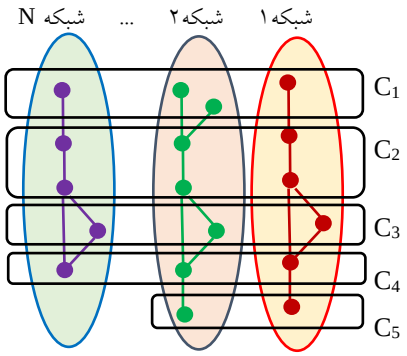
شکل ۲- تطبیق گراف دوبخشی، الف: شبکه‌های ورودی، ب: محاسبه امتیاز نگاشت و ماکزیمم وزن تطبیق دوبخشی

همترازی چندگانه شبکه‌ها: در همترازی چندگانه شبکه‌ها، چندین شبکه (مثلا k شبکه) بعنوان ورودی داده می‌شود، و همترازی آن‌ها به شکل گروه‌بندی پروتئین‌هایی که تکامل یافتند، یا بطور عملکردی در همه شبکه‌ها حفاظت می شوند، اعمال می‌گردد.

یکی از راهکارهای پیشنهادی، همترازی شبکه چندگانه را مانند مسئله خوشه‌بندی در گراف‌های k-بخشی می‌بینند، بطوریکه خوشه‌ها پروتئین‌های مرتبط از نظر عملکرد یا تکامل را با یکدیگر در یک گروه قرار می‌دهند (شکل ۳) (۲۲).

فرض کنید مجموعه k شبکه ورودی PPI بصورت $N = \{N_1, N_2, \dots, N_k | N_i = (V_i, E_i)\}$ باشد. تطبیق‌های ممکن بین پروتئین‌های k شبکه ورودی در گراف k-بخشی $G = (V_1 \cup V_2 \cup \dots \cup V_k, E)$ نمایش داده می‌شود، بطوریکه یال‌ها در E گره‌های شبکه‌های متفاوت را به یکدیگر متصل می‌کنند و وزن مرتبط با آن‌ها بیانگر هزینه تطبیق دو گره است (۲۲).

همان‌طور که در شکل ۳ مشاهده می‌شود همترازی چندگانه شبکه‌ها، همانند مسئله یافتن مجموعه‌ای از خوشه‌های $C = \{C_1, C_2, \dots, C_n\}$ در گراف G است که هر دو معیار شباهت عملکردی بین پروتئین‌های همتراز شده و تعامل حفاظت شده را ماکزیمم نماید. بطوریکه شباهت عملکردی، شباهت درون خوشه‌ای (ICS= Intra-Cluster Similarity) است که مجموع یال‌های وزن‌دار بین پروتئین‌های یک خوشه را بدست می‌آورد؛ و منظور از حفاظت تعامل، حفاظت بین خوشه‌ای (ICC= Inter-Cluster Conservation) است که بیانگر تعداد تعامل‌های حفاظت شده بین پروتئین‌های هر دو خوشه است.



شکل ۳- شبکه چندگانه و خوشه‌بندی گره‌های همتراز شده

بنابراین همترازی چندگانه شبکه‌ها، گراف G را برای ماکزیمم‌سازی خوشه‌بندی C جستجو می‌کند (۲۲):

$$S(C) = \alpha \sum_i ICS(C_i) + (1-\alpha) \sum_{i \neq j} ICC(C_i, C_j) \quad (۲)$$

α پارامتری است بین صفر و یک که تأثیر شباهت درون خوشه‌ای و بین خوشه‌ای را بر روی فرایند همترازی گره متعادل می‌سازد. همترازی سراسری است اگر خوشه‌ها همپوشانی نداشته باشند و مجموعه خوشه‌ها ماکزیمم باشد (یعنی هیچ خوشه اضافی را نتوان به C افزود)؛ در غیر این صورت همترازی موضعی است. همچنین اگر خوشه‌ها در

DOR: 20.1001.1.23832738.1401.35.1.10.4
مقاله پژوهشی

C شامل حداکثر یک پروتئین از هر شبکه باشند نگاشت یک به یک است، در غیر این صورت چند به چند است. هنگامی که α=1 است مسئله همترازی چندگانه شبکه‌ها به تطبیق k–بخشی کاهش می‌یابد و نمی‌تواند راه‌حل بهینه داشته باشد، به همین دلیل تطبیق k–بخشی برای $k \geq 3$، ان پی– سخت است (۲۲).

مواد و روشها

با توجه به وجود روش‌های متفاوت همترازی شبکه، مسئله ارزیابی کیفیت همترازی تولید شده از اهمیت ویژه‌ای برخوردار است. جدول ۱ معیارهای امتیازدهی همترازی استاندارد را در سه گروه همترازی شبکه سراسری، موضعی و چندگانه نشان می‌دهد.

جدول ۱– معیارهای ارزیابی استاندارد

نوع همترازی	نام معیار امتیازدهی	شرح	فرمول محاسبه
موضعی	k-correctness (۲۲)	وجود حداقل k حاشیه‌نویسی مشترک بین دو پروتئین p_1 و p_2 با حاشیه‌نویسی s_1 و s_2	$\lvert s_1 \cap s_2 \rvert \geq k$
	Functional consistency (۲۲)	درصد حاشیه‌نویسی‌های مشترک بین دو پروتئین	$FC(p_1,p_2)=\frac{\lvert s_1 \cap s_2 \rvert}{\lvert s_1 \cup s_2 \rvert}$
	Agreement with reference modules	امتیازدهی بر پایه صحت و فراخوانی مجدد که با استفاده از F-score یکپارچه می‌شود (۲۲).	$F=2\times\frac{P_r \times R_e}{P_r+R_e}$
سراسری	Node correctness	درصد گره‌های همتراز شده‌ای است که بدرستی مطابق با همترازی مرجع نگاشت شدند (۲۰).	
	Node coverage (۲۲)	اندازه‌گیری تعداد گره‌های نرمال شده در بازه [0, 1] نگاشت شده به تعداد گره-های شبکه کوچکتر	$NC(a)=\frac{\lvert V_a \rvert}{\lvert V_1 \rvert}\times 100\%$
	Edge correctness	درصد تعامل‌هایی از شبکه کوچکتر که همتراز با یال‌های شبکه بزرگتر هستند (۲۰).	$EC(a)=\frac{\lvert E_a \rvert}{\lvert E_1 \rvert}\times 100\%$
	Induced conserved sub-structure score	درصد تعامل‌هایی از ناحیه همتراز شده شبکه بزرگتر که با تعامل‌های شبکه کوچکتر همتراز شدند (۳۱).	$ICS(a)=\frac{\lvert E_a \rvert}{\lvert E_{N_2[V_a]} \rvert}\times 100\%$
	Symmetric sub-structure score	تعداد یال‌های همتراز شده به تعداد یال-های شبکه کوچکتر و بزرگتر (۲۲)	$S^3(a)=\frac{\lvert E_a \rvert}{\lvert E_1 \rvert+\lvert E_{N_2[V_a]} \rvert-\lvert E_a \rvert}\times 100\%$
	Size of the largest common connected component	اندازه بزرگترین مؤلفه متصل مشترک (LCC)، LCC بزرگتر یعنی همترازی شامل میزان بزرگتری از ساختار متصل مشترک بین دو شبکه است (۳۱).	
چندگانه	K-coverage	تعداد خوشه‌های همترازی است که شامل پروتئین‌هایی از $k \leq n$ شبکه هستند (۲۲).	
	Exact cluster ratio	درصد همه پروتئین‌هایی که در خوشه‌های واقعی قرار دارند، بیان شود (۲۲).	
	Mean normalized entropy (۲۲)	میانگین آنتروپی نرمال شده (میانگین NE)، p_i کسری از پروتئین‌هایی که در c با عبارت t_i حاشیه‌نویسی شدند و d تعداد عبارات متفاوتی که پروتئین‌ها را در c حاشیه‌نویسی می‌کنند.	$NE(c)=-\frac{1}{\log d}\sum_{i=1}^{d} p_i \times \log p_i$

DOR: 20.1001.1.23832738.1401.35.1.10.4
مقاله پژوهشی

هدف همترازکننده موضعی یافتن مجموعه ماژول‌های همتراز شده متناظر با مسیرهای زیستی یا پیچیدگی‌های پروتئین است؛ لذا بر اساس ارتباط زیستی همترازی‌هایش و بدون در نظر گرفتن ارتباط توپولوژی ارزیابی می‌شود. برخلاف همترازی موضعی، کیفیت همترازی سراسری از روی ارتباط توپولوژی ارزیابی می‌شود، زیرا ایده اصلی این است که نواحی همتراز شده از شبکه‌ها باید الگوهای مشابهی داشته باشند. همترازی چندگانه شبکه‌ها براساس توانایی در تولید خوشه‌هایی که کل شبکه‌های ورودی را پوشش دهد و گروه‌بندی پروتئین‌ها بر اساس شباهت عملکردی، ارزیابی می‌شود.

اغلب تحقیقات در زمینه همترازی شبکه سراسری از معیارهای ارزیابی EC، S^3 و FC استفاده نمودند. لذا در این پژوهش جهت مقایسه کارایی الگوریتم‌های همترازی از معیارهای EC، FC و S^3 استفاده شده است؛ زیرا معیارهای EC و S^3 برای ارزیابی شباهت توپولوژی همترازی، و معیار FC برای ارزیابی شباهت عملکردی همترازی بکار می‌رود. همچنین در حالی که EC بزرگ اجازه می‌دهد یک شبکه کوچک خلوت به یک ناحیه چگال از شبکه بزرگتر نگاشت شود، یک ICS بزرگ اجازه می‌دهد یک ناحیه خلوت از شبکه بزرگتر به یک ناحیه چگال از شبکه کوچکتر نگاشت شود.

امتیاز زیرساختار متقارن (S^3) با مقایسه تعداد یال‌های همتراز شده به تعداد یال‌های شبکه کوچکتر و به تعداد یال‌های ناحیه همتراز شده از شبکه بزرگتر، هر دو شبکه را بررسی می‌کند (۲۲).

پایگاه داده‌های زیادی برای مقایسه کارایی روش‌های همترازی و تجزیه و تحلیل شبکه وجود دارد. پرکاربردترین مجموعه داده، شامل میانکنش‌های پروتئین-پروتئین در گونه‌های مختلف زیستی است؛ و عمومی‌ترین جنبه تحقیقات در این زمینه بر روی تعامل شبکه‌های پروتئین- پروتئین تمرکز دارد. جدول ۲ فهرستی از پرکاربردترین پایگاه داده‌های استفاده شده در تعامل شبکه-های پروتئین-پروتئین را شرح می‌دهد. در این تحقیق از پایگاه داده بیوگرید استفاده شده است.

اغلب تحقیقات در زمینه همترازی شبکه، همانند (۱۶)، (۲۵) و (۳۲) نتایج تجربی خود را بر روی پایگاه داده بیوگرید و پنج گونه زیستی *انسان*، *موش*، *کرم*، *مخمر* و *مگس میوه* تست و بررسی نمودند. برخی تحقیقات نیز همانند (۱۲) و (۱۷) این پنج گونه زیستی را از پایگاه داده InAct و Isobase دریافت نمودند. علت انتخاب این گونه-های زیستی وجود تنوع در میزان پیچیدگی شبکه تعاملی پروتئین‌های این موجودات است.

اعمال و اجرای الگوریتم‌های همترازی شبکه، نیازمند پاکسازی و آماده‌سازی مجموعه داده است. به عنوان مثال مجموعه داده بیوگرید شامل اطلاعات زیادی است. هر سطر بیانگر تعامل بین دو پروتئین است. از میان این اطلاعات، برای همترازی شبکه صرفا به پارامترهای شناسه گره‌ها یا پروتئین‌های موجود در هر سطر نیاز است. امتیاز BLAST یا شباهت توالی موجود بین پروتئین‌ها را باید از روی هستی‌شناسی ژن (Gene Ontology) دریافت نمود (۲۷). سپس مجموعه داده‌ای با سه ستون بدست می‌آید که ستون اول بیانگر شناسه پروتئین اول، ستون دوم بیانگر شناسه پروتئین دوم و ستون سوم بیانگر امتیاز BLAST بین دو پروتئین است. در این بین برخی از محققین، ستون سوم را حذف می‌کنند و در میان کدنویسی امتیاز BLAST را محاسبه می‌نمایند. پایگاه داده Isobase مجموعه داده‌های بیوگرید را آماده استفاده کاربر نموده و امتیاز BLAST را نیز درج نموده است؛ مزیت این پایگاه داده استفاده آسان جهت انجام تست‌های اولیه است و ضعف آن عدم بروزرسانی پس از سال ۲۰۱۴ است، در حالی که دیگر پایگاه داده‌های موجود در جدول ۲ بروزرسانی می‌شوند.

تأثیر الگوریتم‌های فراابتکاری در همترازی شبکه: برای دو شبکه ورودی G_1 و G_2، مسئله یافتن همترازی بین دو

DOR: 20.1001.1.23832738.1401.35.1.10.4
مقاله پژوهشی

شبکه متناظر با جستجو برای یک نگاشت بین گره‌های G_1 و گره‌های G_2 است که تابع هزینه داده شده (کیفیت همترازی) را ماکزیمم می‌سازد.

جدول ۲- فهرست مجموعه داده‌های استاندارد

نام مجموعه داده	قابلیت	دامنه	لینک دسترسی
پایگاه داده شبکه تعامل مولکولی زیستی (BIND)	توصیف کاملی از پیچیدگی‌های مولکولی، تعامل-ها و مسیرهای سلولی	کلیه موجودات	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC165503/ (http://www.bind.ca/)
پایگاه داده تعامل پروتئین‌ها (DIP)	فهرست تعامل‌های آزمایشگاهی	مخمر	http://dip.doe-mbi.ucla.edu/
چارچوب InAct	ذخیره، نمایش، و تحلیل تعامل‌های پروتئین	کلیه موجودات	http://www.ebi.ac.uk/intact
پایگاه داده BioGRID	تعامل‌های پروتئین-پروتئین و ژنتیکی همه اندام‌های اصلی	کلیه موجودات	https://thebiogrid.org/ (https://downloads.thebiogrid.org/)
پایگاه داده تعامل مولکولی (MINT)	ذخیره و بازیابی اطلاعات تجربی	*انسان*، *موش*، *مگس میوه*، مخمر	https://mint.bio.uniroma2.it/
مجموعه داده MPact	تعامل‌های پروتئین-پروتئین برای مخمر	مخمر	http://mips.gsf.de/genre/proj/mpact (https://pubmed.ncbi.nlm.nih.gov/16381906/)
پایگاه داده Isobase	مجموعه‌ای از شبکه‌های PPI یوکاریوت‌ها و مقادیر BLAST	*انسان*، *موش*، *مگس میوه*، *کرم*، مخمر	http://cb.csail.mit.edu/cb/mna/isobase/
مجموعه داده PDB	بانک پروتئین	*انسان*	http://www.wwpdb.org/

برای تولید همترازی، همترازکننده‌ها از یک تابع هدف برای محاسبه امتیاز همترازی استفاده می‌کنند. در حالی که این تابع در همترازکننده‌ها متفاوت است اما می‌توان آن را به شکل کلی زیر نوشت (۲۲):

$$S(f) = \sum_{u \in V_1} n\big(u, f(u)\big) + \sum_{(u,v) \in E_1} e(u, f(u), v, f(v)) \quad (۳)$$

DOR: 20.1001.1.23832738.1401.35.1.10.4
مقاله پژوهشی

بطوریکه $n: V_1 \times V_2 \rightarrow R^+$ امتیاز نگاشت یک گره از V_1 به یک گره در V_2 است، و $e: E_1 \times E_2 \rightarrow R^+$ امتیاز نگاشت یک یال از E_1 به یک یال از E_2 است.

در مسئله همترازی اندازه فضای جستجو بزرگ است، زیرا همه نگاشت‌های بین گره‌های شبکه‌های مورد مقایسه را در برمی‌گیرد. عدم سازگاری محاسباتی مسئله که ناشی از «ان پی-کامل» بودن مسئله زیرگراف‌های ایزومورف است، باعث می‌شود برای حل آن نیاز به توسعه روش‌های فراابتکاری (نظریه‌های تقریبی) داشته باشیم.

بین همترازی موضعی و سراسری ارتباط واضحی وجود دارد. برای مثال، هدف هر دو یافتن شباهت‌های توپولوژی و عملکردی بین شبکه‌های مورد مقایسه است تا دانش زیستی گونه‌های مورد مطالعه را به گونه‌هایی که کمتر مطالعه شده‌اند تعمیم دهند (۱۳). با این حال، محققان در این دو زیرمجموعه الگوریتم‌های مستقلی ارائه نمودند. در این راستا می‌توان به الگوریتم‌های همترازی موضعی PathBLAST (۱۹) و AlignMCL (۲۸)؛ الگوریتم‌های همترازی سراسری مانند IsoRank (۳۳)، MAGNA (۳۲) و SANA (۱۶، ۲۵) اشاره نمود. الگوریتم PathBLAST به کشف مسیرهای سلولی بر اساس امتیاز BLAST یا شباهت توالی می‌پردازد و شباهت توپولوژی را نیز در نظر می‌گیرد. الگوریتم AlignMCL با استفاده از خوشه‌بندی مارکوف سعی در یافتن همترازی‌های موضعی دارد. الگوریتم IsoRank بر پایه شباهت توپولوژی و محاسبه امتیاز PageRank عمل می‌کند. به همین ترتیب هر یک از روش متفاوتی بهره می‌برند؛ در نتیجه، بسیاری از الگوریتم‌های همترازی موضعی و بسیاری از الگوریتم‌های همترازی سراسری وجود دارد که بر فرضیه‌های متفاوت تکیه می‌کنند و از نظریه‌های متفاوت استفاده می‌کنند تا توابع هزینه‌های متفاوتی را به حداکثر برسانند.

بعنوان مثال، بسیاری از الگوریتم‌ها سعی می‌کنند برخی از توابع هزینه را بطور عمده بر اساس توپولوژی، بهینه‌سازی کنند، در حالی که بسیاری دیگر به منظور افزایش ارتباطات کاربردی همترازی‌ها طراحی شده‌اند. بنابراین، مقایسه مستقیم یک روش همترازی موضعی و یک روش همترازی سراسری، بی‌اهمیت است؛ بلکه هر یک با روش‌های دسته خود مقایسه می‌شوند. بنابراین پرسش این است که کدام یک از آن‌ها را استفاده کنیم: همترازی موضعی، همترازی سراسری یا نظریه ترکیبی که با هر دو سازگار باشد. سازگاری ممکن بین این دو جنبه همترازی شبکه یک مسئله باز برای تحقیق است که باید در آینده بطور عمیق بررسی شود (۱۳).

اهمیت روش‌های تطبیق گراف، تطبیق شبکه و همترازی شبکه بیانگر این حقیقت است که این پدیده می‌تواند مفاهیم ریاضی را به شکلی که امروزه علم شبکه نامیده می‌شود نمایش دهد. با بیان تفاوت‌های کمی در شبکه‌ها بکارگیری بسیاری از استانداردهای روش‌های آماری و یادگیری ماشین امکان‌پذیر است (۱۰).

در این پژوهش الگوریتم‌های همترازی شبکه‌های زیستی را بر اساس نوع همترازی به چهار دسته کلی همترازی سراسری، همترازی موضعی، همترازی چندگانه شبکه‌ها، و یکپارچه‌سازی همترازی تقسیم شده‌اند.

در میان الگوریتم‌های همترازی سراسری، MAGNA (۳۲)، MAGNA++ (۳۵) و الگوریتم برنامه‌سازی پویای DynaMAGNA++ (۳۶) برای همترازی شبکه از الگوریتم ژنتیک استفاده می‌کنند.

همچنین روش MeAlign (۱۲) و OptnetAlign (۸) از الگوریتم میمتیک استفاده می‌کنند که ترکیبی از الگوریتم ژنتیک با یک پالایش جستجوی محلی است. الگوریتم جستجوی SANA همترازی شبکه را با استفاده از الگوریتم تبرید شبیه‌سازی شده انجام می‌دهد (۲۵)، (۱۶). همترازکننده PSONA مبتنی بر الگوریتم فراابتکاری بهینه-سازی توده ذرات است (۱۷).

DOR: 20.1001.1.23832738.1401.35.1.10.4
مقاله پژوهشی

الگوریتم‌های همترازی موضعی شبکه، برای همترازی از استراتژی انتخاب و توسعه (۴)، (۳۱)؛ فاصله ویرایشی گراف (۲۰)، روش‌های فرابتکاری بهمراه روش‌های حریصانه (Greedy methods) (۷)، (۹)، (۱۴)، (۲۳)، (۳۰)، (۳۴) استفاده می‌کنند.

روش همترازی سراسری GEDEVO-M (۱۸) بر پایه فاصله ویرایشی گراف ((Graph Edit Distance (GED عمل می‌کند. این روش با استفاده از الگوریتم ژنتیک بدنبال همترازی با کمترین مجموع GED از بین همه جفت همترازی‌های شبکه است.

روش MultiMAGNA++ (۳۷) برای همترازی شبکه چندگانه سراسری ارائه شده و توسعه‌یافته‌ی روش MAGNA++ است. این روش از الگوریتم ژنتیک برای تولید همترازی‌های شبکه چندگانه یک به یک استفاده می‌کند که یال حفاظت شده بین شبکه‌های همتراز شده را ماکزیمم می‌سازد.

الگوریتم ACOTS-MGA برای تطبیق سایت‌های پیوند پروتئینی ارائه شده است که بر اساس الگوی میمتیک و با استفاده از روش بهینه‌سازی کلونی مورچگان (ACO) مجموعه‌ای راه‌حل را ایجاد می‌کند، سپس بهترین راه‌حل برای پیاده‌سازی را با جستجوی ممنوعه انتخاب می‌کند تا کیفیت راه‌حل را بهبود دهد (۱۵).

همان‌طور که همترازی موضعی شبکه کیفیت عملکردی بالا و کیفیت همترازی توپولوژیکی پایینی دارد، همترازی سراسری شبکه نیز کیفیت توپولوژیکی بالا و کیفیت همترازی عملکردی پایینی دارد. نظریه IGLOO (۲۶) مؤلفه‌های الگوریتمی همترازی موضعی و سراسری شبکه را با امید به ترکیب دو نوع همترازی شبکه، یکپارچه می‌سازد. ابزار Ulign (۲۴) هشت روش همترازی را بطور همزمان برای نگاشت و انتقال در میان همه پروتئین‌های شبکه‌های PPI استفاده می‌کند، تا بهترین همترازی را انتخاب کند. همترازکننده محلی سراسری GLAlign روشی است که کارایی همترازکننده‌های شبکه محلی را با بهره‌گیری از همترازی سراسری بهبود می‌بخشد (۲۹).

نتایج

نتایج حاصل از اجرا نشان داد که روش‌های همترازی شبکه‌ای که از الگوریتم‌های فرابتکاری استفاده می‌کنند دارای ثبات رفتاری بیشتری نسبت به سایر روش‌ها هستند. بدین معنی که با بزرگتر شدن و یا پیچیده‌تر شدن ساختار شبکه میزان کیفیت همترازی کاهش چندانی ندارد. این امر با بررسی عملکرد الگوریتم‌های همترازی شبکه و محاسبه معیارهای ارزیابی بدست آمد که در ادامه شرح آن و نحوه مقایسه آورده شده است.

در این تحقیق برای مقایسه عملکرد الگوریتم‌های فرابتکاری با دیگر روش‌های همترازی ۱۵ الگوریتم معروف و پرکاربرد NATALIE (۹)، GHOST (۳۱)، SPINAL (۴)، NETAL (۳۰)، PISWAP (۷)، HubAlign (۱۴)، L-GRAAL (۲۳)، OptNetAlign (۸)، CytoGEDEVO (۲۱)، MAGNA++ (۳۵)، WAVE (۳۴)، MeAlign (۱۲)، SANA (۱۶، ۲۵)، PSONA (۱۷) و ACOTS-MGA (۱۵) انتخاب شده است.

از میان مجموعه داده‌های سطح زیستی (۶) نسخه BIOGRID-3.5.170 (۵) برگزیده شد که مشخصات هر یک از داده‌های منتخب مربوط به *انسان*، *موش*، *کرم*، *مگس میوه* و *مخمر* در جدول ۳ آورده شده است.

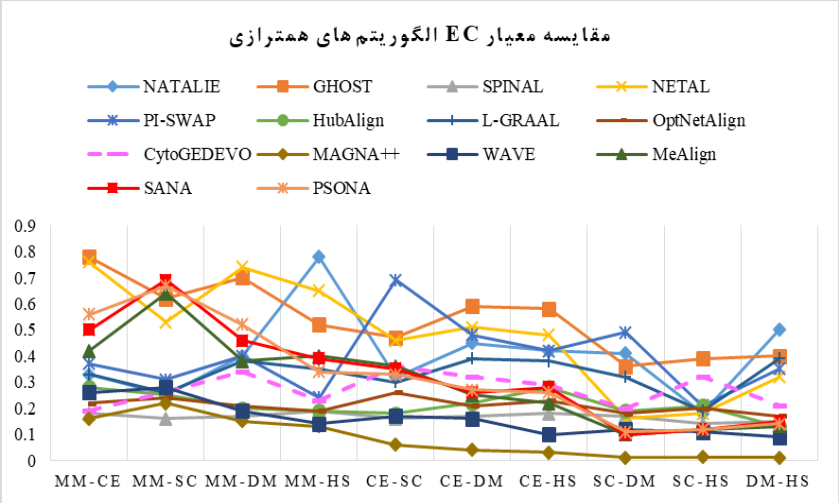
جدول ۳- ویژگی‌های شبکه PPI از پنج گونه

نام گونه	نام مجموعه داده	تعداد گره‌ها	تعداد یال‌ها
انسان	*Hsapi(HS)*	۹۶۳۳	۳۴۳۲۷
موش	*Mmusc(MM)*	۲۹۰	۲۴۲
کرم	*Celeg(CE)*	۲۸۰۵	۴۴۹۵
مگس میوه	*Dmela(DM)*	۷۵۱۸	۲۵۶۳۵
مخمر	*Scere(SC)*	۵۴۹۹	۳۱۲۶۱

DOR: 20.1001.1.23832738.1401.35.1.10.4
مقاله پژوهشی

دلیل انتخاب این پنج گونه، کاربرد بسیار زیاد آن‌ها در اثبات کارایی الگوریتم‌های همترازی شبکه است. شکل‌های ۴ تا ۶ متناظراً معیارهای ارزیابی EC، FC و S^3 الگوریتم-های مختلف را بر روی این پنج گونه نشان می‌دهند.

در شکل ۴ مشاهده می‌شود که تقریبا در تمام الگوریتم‌ها میزان درستی تشخیص یال یا صحت یال (EC) با بزرگ شدن شبکه‌ها و افزایش پیچیدگی آن‌ها کاهش می‌یابد.

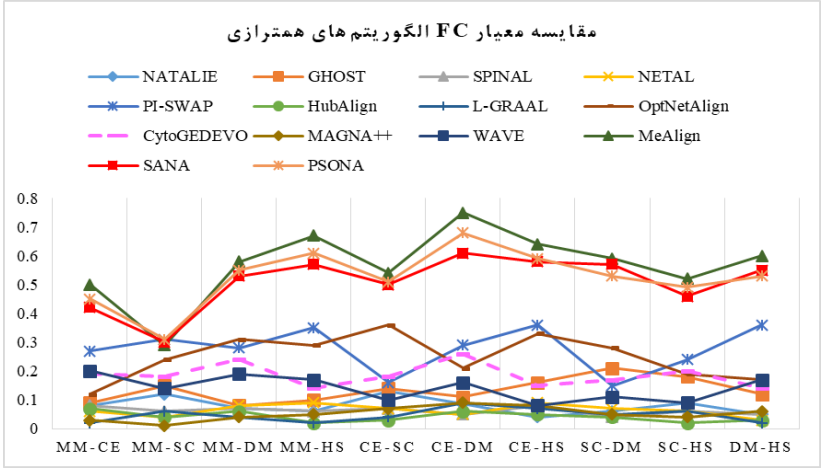


شکل ۴- مقایسه معیار EC الگوریتم‌های همترازی

بدیهی است تشخیص صحیح یال در گراف همترازی بمنزله تشخیص همترازی درست بین دو پروتئین یا گره متناظر در گراف است؛ لذا هر چقدر مقدار EC بیشتر باشد الگوریتم بهتر عمل کرده است.

بنابر شکل ۴، الگوریتم GHOST که از استراتژی انتخاب و توسعه بهره می‌برد دارای ثبات رفتاری بهتری نسبت به سایر الگوریتم‌ها است. الگوریتم NETAL که از روش مکاشفه‌ای استفاده می‌کند در رتبه دوم قرار دارد؛ و پس از آن می‌توان به الگوریتم‌های PSONA و SANA که جزء روش‌های فراابتکاری هستند اشاره نمود.

همان‌طور که در شکل ۵ مشاهده می‌شود میزان تشخیص صحیح پروتئین‌هایی با عملکرد مشابه یعنی معیار FC یا شباهت عملکردی الگوریتم‌های مختلف با یکدیگر متفاوت است. در این میان، الگوریتم‌های MeAlign، PSONA و SANA از بهترین قدرت تشخیص شباهت عملکردی بهره می‌برند که همگی جزء روش‌های فراابتکاری هستند.

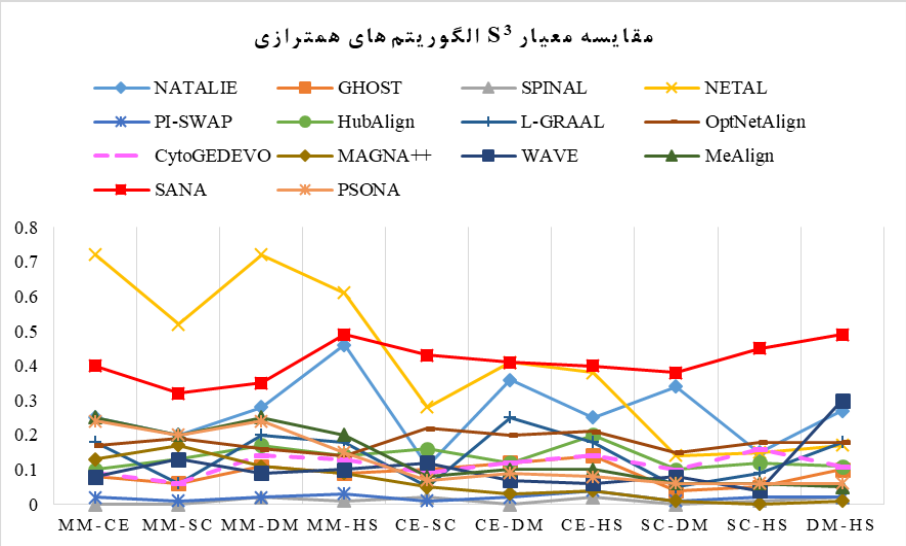


شکل ۵- مقایسه معیار FC الگوریتم‌های همترازی

DOR: 20.1001.1.23832738.1401.35.1.10.4
مقاله پژوهشی

شکل ۶ معیار S^3 الگوریتم‌های همترازی را مقایسه می‌کند. همان‌گونه که ملاحظه می‌شود در میان الگوریتم‌های مختلف، روش SANA از بیشترین پایداری و ثبات مقدار بهره می‌برد. پس از آن می‌توان به الگوریتم NETAL اشاره کرد که در برخی جفت گونه‌ها مانند MM-CE و MM-DM بسیار خوب عمل کرده است. لازم به ذکر است که هر دو الگوریتم SANA و NETAL از روش فراابتکاری و مکاشفه‌ای استفاده می‌کنند.

همان‌طور که در شکل‌های ۴ تا ۶ مشاهده می‌شود هر یک از الگوریتم‌ها بر روی یک یا چند مجموعه داده بهتر عمل می‌کنند. الگوریتم‌های فراابتکاری از جمله روش‌های همترازی هستند که در اغلب موارد هر سه معیار EC، FC و S^3 را ماکزیمم نمودند و این نشان از کارایی روش‌های فراابتکاری است.

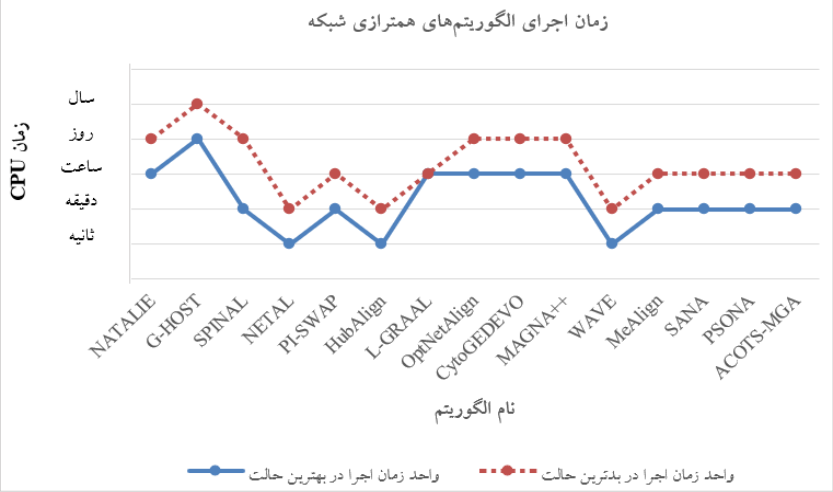


شکل ۶- مقایسه معیار S^3 الگوریتم‌های همترازی

بحث

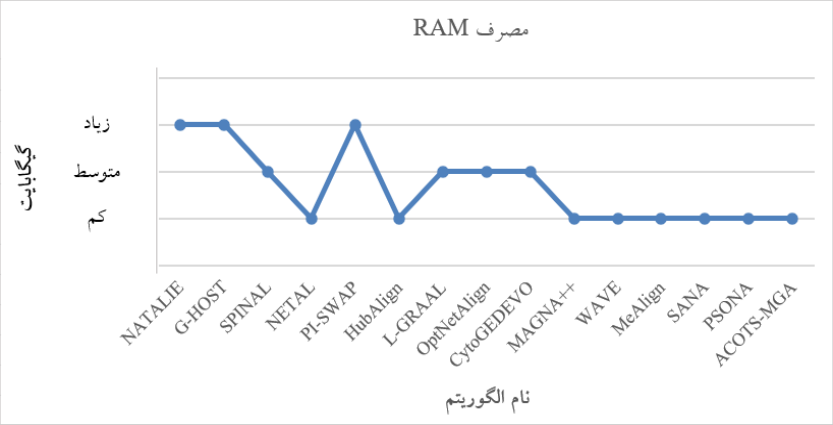
همان‌طور که ذکر شد یکی از چالش‌ها در تحقیقات زیستی، بکارگیری داده‌های زیستی در حجم بالاست که نیازمند استفاده از روش‌های محاسباتی و الگوریتم‌های هوشمند است (۳). یکی از مجموعه داده‌های حجیم، داده‌های حاصل از میانکنش پروتئین‌ها است که نقش اساسی در فرآیندهای سلولی دارند (۱). می‌دانیم که روش-های محاسباتی جهت بررسی میانکنش دو پروتئین نیازمند استفاده از ساختار سه بعدی دو پروتئین هستند (۲). اما در بخش قبل نشان داده شد که الگوریتم‌های فراابتکاری با استفاده از ساختار شبکه‌های زیستی، می‌توانند نقش مؤثری در کشف همترازی‌های موجود بین دو شبکه PPI داشته باشند. در این بخش جهت مقایسه بهتر، زمان اجرا و میزان مصرف حافظه الگوریتم‌ها نیز با یکدیگر مقایسه شده است. شکل ۷ زمان اجرای الگوریتم‌های همترازی را برحسب ثانیه، دقیقه، ساعت، روز و سال نشان می‌دهد. بر اساس این مقایسه می‌توان مشاهده کرد که الگوریتم‌های NETAL، HubAlign و WAVE از کمترین زمان اجرا بهره می‌برند که از روش‌های مکاشفه‌ای استفاده نمودند و GHOST از بیشترین زمان اجرا بهره می‌برد.

DOR: 20.1001.1.23832738.1401.35.1.10.4
مقاله پژوهشی



شکل ۷- مقایسه زمان اجرای الگوریتم‌های همترازی

لیکن این نتیجه به میزان پیچیدگی و تعداد تعامل پروتئین‌ها در شبکه‌های مورد بررسی وابسته است؛ لذا در شکل ۷ میزان پیچیدگی شبکه‌های هر الگوریتم بر اساس حداقل (بهترین حالت) و حداکثر (بدترین حالت) تعداد یال‌ها و گره‌ها آورده شده است. آنچه در شکل ۷ حائز اهمیت می‌باشد این است که بدترین زمان اجرای الگوریتم‌های همترازی سراسری MeAlign، SANA، PSONA و ACOTS-MGA که همگی جزء روش‌های فراابتکاری هستند برحسب دقیقه است؛ در حالی که بدترین زمان اجرای سایر الگوریتم‌ها بر حسب ساعت و روز هستند؛ این یعنی الگوریتم‌های فراابتکاری از نظر زمان اجرا نیز بهتر و سریع‌تر از سایر الگوریتم‌ها عمل می‌کنند. شکل ۸ میزان مصرف حافظه RAM هر الگوریتم را نشان می‌دهد.

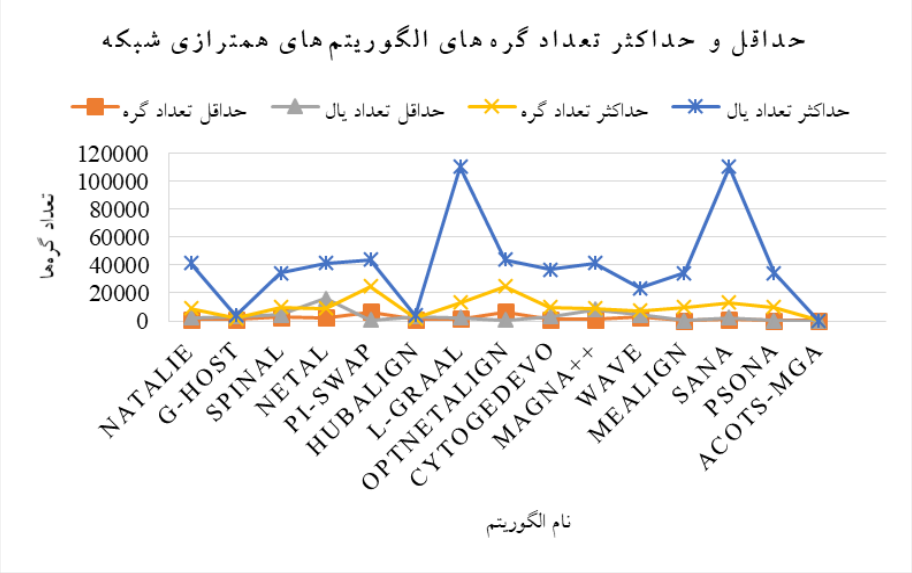


شکل ۸- مقایسه مصرف حافظه الگوریتم‌های همترازی

با توجه به شکل ۸ کمترین میزان مصرف حافظه به الگوریتم‌های NETAL، HubAlign، MAGNA++، WAVE، MeAlign، SANA، PSONA و ACOTS-MGA برمی‌گردد که اغلب آن‌ها از الگوریتم‌های فراابتکاری بهره می‌برند. در شکل ۸ بیشترین میزان مصرف حافظه مربوط به الگوریتم‌های NATALIE، GHOST و PI-SWAP است که از روش‌های ریاضی و حریصانه استفاده می‌کنند.

در شکل ۹ حداقل و حداکثر تعداد یال‌ها و گره‌هایی که محققین در الگوریتم‌های همترازی خود استفاده کردند، نمایش داده شده است.

DOR: 20.1001.1.23832738.1401.35.1.10.4
مقاله پژوهشی



شکل ۹- مقایسه حداقل و حداکثر تعداد گره‌ها و یال‌ها

بر اساس شکل ۹ می‌توان نتیجه گرفت که الگوریتم‌هایی مانند OptNetAlign و PI-SWAP بیشترین تعداد گره (۲۴۸۵۵ گره) و الگوریتم‌های MeAlign و PSONA شبکه‌هایی با حداقل تعداد گره (۲۹۰ گره) و حداقل تعداد یال (۲۴۲ یال) را بررسی می‌نمایند.

جدول ۴- مقایسه مرتبه اجرای الگوریتم‌های همترازی

نام الگوریتم	مرتبه اجرایی
PSONA	$O(NMV\,log(\|V\|))$
DynaMAGNA+ +	$O(N(p\|V\| + p\|E\|\,\log(\|E\|) + p\log(p)))$
MAGNA++	$O(\|E1\| + \|E2\|)$
MAGNA	$M\|V\|)\ O(N(M\|E\|+log(\|E\|)+ + Mlog(M))$
HubAlign	$O(\|V\|^2log\|V\|)$
MI-GRAAL	$O(\|V\|^2log\|V\|)$
NETAL	$O(\|V\|^2log^2\|V\|)$
IsoRank	$O(\|V\|^4)$

الگوریتم‌های SANA و L-GRAAL نیز شبکه‌هایی با بیشترین تعداد یال (۱۱۰۵۲۸ یال) را بررسی می‌کنند. در این میان الگوریتم ACOTS-MGA حداقل ۴ گراف و حداکثر ۳۲ گراف را همزمان در نظر می‌گیرد.

مرتبه اجرایی برخی از الگوریتم‌های همترازی در جدول۴ نشان داده شده است. در جدول ۴، p اندازه جمعیت، N تعداد نسل‌ها، V مجموعه رئوس، E مجموعه یال‌ها، و M تعداد تکرار می‌باشد.

بنابر مطالعات انجام شده می‌توان الگوریتم‌های موجود در شبکه همترازی سراسری (GNA= Global Network Alignment) و شبکه همترازی موضعی (LNA= Local Network Alignment) را از دیدگاه سیستماتیک بررسی نمود. اگرچه هدف LNA یافتن ماژول‌های حفاظت شده با عملکرد بالاست، ماژول‌های روش‌های LNA موجود می‌توانند بخش کوچکی را پیدا کنند. به عبارت دیگر در حالی که هدف GNA یافتن کل نواحی شباهت توپولوژیکی بین شبکه‌های مورد مقایسه است، نواحی که روش‌های GNA موجود می‌توانند پیدا کنند معمولاً از نظر عملکرد حفاظتی ضعیف هستند (۱۳).

از این رو محققین بررسی کردند که آیا ممکن است هر دو روش، نواحی حفاظت شده با توپولوژی بزرگ و عملکرد بالا از شباهت شبکه را پیدا کنند، که نه تنها اهداف کاربردی/ طراحی هر یک از روش‌های LNA و GNA را برآورده کند بلکه بر معایب آن‌ها نیز چیره شود (۱۳)، (۲۶).

DOR: 20.1001.1.23832738.1401.35.1.10.4
مقاله پژوهشی

آن‌ها شواهدی ارائه کردند که: ۱) ماژول‌های کوچک حفاظت شده با عملکرد بالا از نواحی LNA موجود می‌توانند گسترش یابند تا کیفیت توپولوژیکی خود را بدون کاهش کیفیت عملکردی‌شان بهبود بخشند. ۲) نواحی بزرگ شباهت توپولوژیکی از روش‌های GNA موجود بطور معمول کارایی معناداری ندارند اما می‌توانند به نواحی معنادار کاربردی معطوف شوند، که بدون کاهش کیفیت توپولوژیکی کیفیت عملکردی را بهبود می‌بخشند.

چنین یکپارچه‌سازی کارآمدی از LNA و GNA که هدف اصلی هر یک از آن‌ها را همزمان حفظ می‌کند، هر دو همترازی موضعی و سراسری را بهبود می‌بخشد. از دیدگاه محققین هدف برد–برد به این صورت تعریف می‌شود که نواحی شبکه موضعی حفاظت شده عملکردی را پیدا نمایند و هدف آن‌ها یافتن نواحی شبکه سراسری حفاظت شده با امتیاز توپولوژیکی بالا است (۱۳). تعادل بین اندازه ناحیه شبکه همتراز شده (کیفیت توپولوژیکی) مهم است و این کیفیت عملکردی ممکن است نیاز به تنظیم پارامترهای داده شده روش یکپارچه‌سازی LNA-GNA داشته باشد.

علاوه بر ایده فوق چندین سؤال دیگر از دیدگاه عملی باقی می‌ماند که بصورت زیر است: کدام روش را دانشمندان زیست‌شناسی استفاده می‌کنند؟ طبق ادبیات تحقیق می‌توان گفت در حالی که LNA بطور کلی نمی تواند نواحی بزرگ مشابه را پوشش دهد، GNA قادر به بازیابی مسیرهای معنادار زیستی یا نمونه‌های پروتئینی حفاظت شده تکاملی است. لذا اغلب دانشمندان همترازی سراسری را استفاده می‌کنند. علاوه بر این، پاسخ سؤال فوق با ایده‌های دیگر نیز مرتبط است: چگونه می‌توان دو همترازی متفاوت را مقایسه نمود؟ و چگونه کیفیت همترازی را اندازه‌گیری نماییم؟ بر اساس تحقیقات، بخشی از این مسئله حل شده است و تا کنون معیارهای ارزیابی مختلفی تعریف شده است (جدول ۱). بنابراین برای اندازه‌گیری کیفیت همترازی می‌توان با توجه به نوع همترازی از معیارهای موجود در جدول ۱ بهره برد.

برخی محققین روش‌های GNA را بررسی کردند و معیار جدیدی برای کیفیت همترازی ارائه دادند. برخی از آن‌ها روش‌های LNA را بررسی نمودند و چارچوبی برای بهبود پایداری همترازی‌های موضعی ایجاد کردند. اخیراً برخی تحقیقات در مورد یکپارچه‌سازی LNA و GNA بررسی کردند و روشی برای مقایسه بین آن‌ها پیشنهاد دادند، همانند (۲۴)، (۲۶) و (۲۹). لازم به ذکر است که مقایسه حتی در میان همترازی‌های با نوع یکسان (یعنی LNA یا GNA) نیز پیچیده است، و بطور خاص مقایسه همترازی-هایی از کلاس‌های متفاوت (یعنی LNA و GNA) سخت-تر است. بنابراین محققین معیارهای جدیدی از کیفیت همترازی ارائه دادند که بر روی هر دو GNA و LNA کار می‌کند. لذا می‌توان گفت جهت مقایسه دو همترازی با انواع متفاوت بهتر است از روش‌های یکپارچه‌سازی (۲۴)، (۲۶)، و (۲۹) بهره برد؛ و در صورت عدم امکان یکپارچه-سازی برای مقایسه کیفیت از معیارهای ارزیابی که در هر دو نوع همترازی، قابل محاسبه باشند، استفاده نمود.

با گسترش تحقیقات و رشد روزافزون داده‌های زیستی، توسعه همترازکننده‌ای مبتنی بر محاسبات ابری (Cloud Computing) که از فناوری داده‌های حجیم (Big Data) بهره ببرد؛ می‌تواند جزء کارهای آتی محققین قرار گیرد.

نتیجه‌گیری

بطور کلی نتایج آزمایشی و تحلیل‌های نظری نشان می‌دهند که استفاده از الگوریتم‌های فراابتکاری از جمله الگوریتم-های ژنتیک، میمتیک، بهینه‌سازی توده ذرات، تبرید شبیه-سازی شده و کلونی مورچگان نقش مؤثری در بهینه‌سازی معیارهای ارزیابی همترازی شبکه، کاهش زمان اجرا و مصرف حافظه دارد. در این رابطه، ایجاد روش‌هایی برای یکپارچه‌سازی همترازی سراسری و محلی، و ایجاد معیارهای استاندارد برای مقایسه عملکرد و کیفیت همترازی روش‌های یکپارچه‌سازی، همچنان از زمینه‌های

DOR: 20.1001.1.23832738.1401.35.1.10.4
مقاله پژوهشی

تحقیقاتی باز و فعالی در حوزه همترازی شبکه به حساب می‌آیند.

منابع

۱- پورشیخ ع، اصغری م، عبدالمالکی پ. (۱۳۹۴). پیشگویی عملکرد اتّصال پروتئینها به ریبونوکلئیک اسید بر اساس خواص فیزیکوشیمیایی آنها به کمک روش لوژستیک رگرسیون، *مجله پژوهش‌های سلولی و مولکولی* ۲۸(۱)، ۴۵-۵۳.

۲- رضوان‌نژاد ا، لطفی ص، بوستان آ. (۱۳۹۸). مدل‌سازی پروتئین استرس گرمایی ۷۰ (HSP70) زنبور عسل به روش همولوژی مدلینگ و شبیه‌سازی ملکولی و اتصال آن به HSP40. *مجله پژوهش‌های سلولی و مولکولی* ۳۲(۴)، ۴۸۲-۴۹۱.

۳- مشیری م، قادری‌زفره‌ای م، قانع گلمحمدی ف. (۱۳۹۴). مقایسه الگوریتم‌های برپایه یادگیری ماشین بر دقت تخمین داده‌های گمشده حاصل از آزمایش‌های ریزآرایه. *مجله پژوهش‌های سلولی و مولکولی* ۲۸(۴)، ۶۱۲-۶۲۲.

4- Aladağ, Ahmet E., and Cesim Erten. "SPINAL: scalable protein interaction network alignment." Bioinformatics 29, no. 7 (2013): 917-924.

5- Biogrid download file repository, https://downloads.thebiogrid.org/BioGRID/Release-Archive/BIOGRID-3.5.170/, Available on: March 2019.

6- Chatr-Aryamontri A, et al. (2013) The biogrid interaction database: 2013 update. Nucleic Acids Res., 41, D816–D823.

7- Chindelevitch L, Cheng-Yu M, Chung-Shou L, and Bonnie B. "Optimizing a global alignment of protein interaction networks." Bioinformatics 29, no. 21 (2013): 2765-2773.

8- Clark, Connor, and Jugal Kalita. "A multiobjective memetic algorithm for PPI network alignment." Bioinformatics 31, no. 12 (2015): 1988-1998.

9- El-Kebir M, Jaap H, and Gunnar W. K. "Lagrangian relaxation applied to sparse global network alignment." In IAPR International Conference on Pattern Recognition in Bioinformatics, pp. 225-236. Springer, Berlin, Heidelberg, 2011.

10- Emmert-Streib F, Matthias D, and Yongtang S. "Fifty years of graph matching, network alignment and network comparison." Information Sciences 346 (2016): 180-197.

11- Faisal F. E, Han Z, and Tijana M. "Global network alignment in the context of aging." IEEE/ACM Transactions on Computational Biology and Bioinformatics 12, no.1 (2014): 40-52.

12- Gong M, Zhenglin P, Lijia M, and Jiaxiang H. "Global biological network alignment by using efficient memetic algorithm." IEEE/ACM transactions on computational biology and bioinformatics 13, no. 6 (2016): 1117-1129.

13- Guzzi, P. H, and Tijana M. "Survey of local and global biological network alignment: the need to reconcile the two sides of the same coin." Briefings in bioinformatics 19, no. 3 (2017): 472-481.

14- Hashemifar S, and Jinbo X. "HubAlign: an accurate and efficient method for global alignment of protein–protein interaction networks." Bioinformatics 30, no. 17 (2014): i438-i444.

15- Ha T. N, and Hoang X. H. "A new memetic algorithm for multiple graph alignment." VNU Journal of Science: Computer Science and Communication Engineering 34, no. 1 (2018).

16- Hayes WB. An introductory guide to aligning networks using SANA, the Simulated Annealing Network Aligner. In Protein-Protein Interaction Networks 2020 (pp. 263-284). Humana, New York, NY.

17- Huang J, Maoguo G, and Lijia M. "A global network alignment method using discrete particle swarm optimization." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 15, no. 3 (2018): 705-718.

18- Ibragimov R, Malek M, Baumbach J, Guo J. "Multiple graph edit distance: Simultaneous topological alignment of multiple protein-protein interaction networks with an evolutionary algorithm." In Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation. ACM, (2014): 277–284.

19- Kelley BP, Sharan R, Karp RM, et al. "Conserved pathways within bacteria and yeast as revealed by global protein network alignment." Proceedings of the National Academy of Sciences, 2003; 100 (20): 11394–11399.

20- Kuchaiev O, Milenkovi´c T, Memi˘sevi´c V, Hayes W, Pr˘zulj N. Topological network

DOR: 20.1001.1.23832738.1401.35.1.10.4
مقاله پژوهشی

alignment uncovers biological function and phylogeny. Journal of the Royal Society Interface, 2010;7(50):1341–1354.

21- Malek, M, Ibragimov R, Albrecht M, and Baumbach J. "CytoGEDEVO—global alignment of biological networks with Cytoscape." Bioinformatics 32, no. 8 (2015): 1259-1261.

22- Malod-Dognin N, and Prˇzulj N. "9 Network Alignment." Analyzing Network Data in Biology and Medicine: An Interdisciplinary Textbook for Biological, Medical and Computational Scientists (2019): 369.

23- Malod-Dognin N, and Pržulj N. "L-GRAAL: Lagrangian graphlet-based network aligner." Bioinformatics 31, no. 13 (2015): 2182-2189.

24- Malod-Dognin. N, Ban. K, and Pržulj. N. "Unified alignment of protein-protein interaction networks." Scientific reports 7, no. 1 (2017): 953.

25- Mamano N, Wayne B. H. "SANA: simulated annealing far outperforms many other search algorithms for biological network alignment." Bioinformatics 33, no. 14 (2017): 2156-2164.

26- Meng L, Joseph C, Aaron S, and Tijana M. "IGLOO: Integrating global and local biological network alignment." arXiv preprint arXiv:1604.06111 (2016).

27- Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. "GO Enrichment Analysis: PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools." Nucleic Acids Res. Jan 2019;47(D1):D419-D426.

28- Mina M, Guzzi PH. "AlignMCL: Comparative analysis of protein interaction networks through Markov clustering. " In 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW). IEEE;2012. pp. 174–181.

29- Milano. M, Guzzi P. H., and Cannataro.M. "GLAlign: A Novel Algorithm for Local Network Alignment." IEEE/ACM transactions on computational biology and bioinformatics (2018).

30- Neyshabur, B, Khadem A, Hashemifar S, and Arab S.S. "NETAL: a new graph-based method for global alignment of protein–protein interaction networks." Bioinformatics 29, no. 13 (2013): 1654-1662.

31- Patro R, Kingsford C. "Global network alignment using multiscale spectral signatures". Bioinformatics, 2012;28(23):3105–3114.

32- Saraph V, Milenković T. "MAGNA: maximizing accuracy in global network alignment." Bioinformatics 30, no. 20 (2014): 2931-2940.

33- Singh, Rohit, Jinbo Xu, and Bonnie Berger. "Global alignment of multiple protein interaction networks with application to functional orthology detection." Proceedings of the National Academy of Sciences 105, no. 35 (2008): 12763-12768.

34- Yihan S, Crawford J, Tang J, and MilenkovićT,"Simultaneous optimization of both node and edge conservation in network alignment via WAVE." In International Workshop on Algorithms in Bioinformatics, pp. 16-39. Springer, Berlin, Heidelberg, 2015.

35- Vijayan V, Saraph V, and Milenković T. "MAGNA++: Maximizing Accuracy in Global Network Alignment via both node and edge conservation." Bioinformatics 31, no. 14 (2015): 2409-2411.

36- Vijayan V, Dominic C, and Milenković T. "Alignment of dynamic networks." Bioinformatics 33, no. 14 (2017): i180-i189.

37- Vijayan V, Milenkovic T. "Multiple network alignment via multiMAGNA++." In Proceedings of the 15th International Workshop on Data Mining in Bioinformatics (BIOKDD) at the 22nd ACM SIGKDD 2016 Conference on Knowledge Discovery & Data Mining (KDD); ACM SIGKDD (2016).

مقاله پژوهشی
DOR: 20.1001.1.23832738.1401.35.1.10.4

Effects of Meta-Heuristic Algorithms in Protein-Protein Interaction Networks Alignment in Five Biological Species

Mahdipour E. and Ghasemzadeh M.

Faculty of Computer Engineering, Yazd University, Yazd, I.R. of Iran.

Abstract

The biological knowledge of different species can be transferred to the conserved sequence regions via genomic sequence alignment. Similarly, through biological network alignment, knowledge of the conserved regions of molecular networks can be transferred to different conserved regions of different species. Therefore, relying on biological network alignment, we can extend the traditional "sequence-based homology" concept to the new concept of "network-based homology". Discovery of networks alignment is especially important because of its applications, such as discovering new drugs, tracking disease progression, or predicting users' behavior on social networks. In this regard, the main challenge is that the problem of finding the alignments in two graphs is NP-hard. In situations like this, we can use the relatively fast meta-heuristic algorithms to find some acceptable approximate solutions. The main contribution of this research consists of conducting a comparison over the network alignment algorithms, based on the respective evaluation criteria, execution time, memory consumption and complexity of the testing networks. The experimental results are obtained from running the relevant algorithms on the well-known BioGRID dataset. Evaluations indicate that among other methods, using genetic algorithm, memetic, particle swarm optimization, simulated annealing and the ant colony, could yield more valuable results. The named methods apply appropriate heuristics to generate and investigate only a very small subset of the whole search space with the highest probability of holding a solution; therefore, they often can find the optimal solution or some acceptable solutions in a relatively short time.

Key words: Graph matching, Isomorphic subgraph, Meta-heuristic algorithm, Network alignment, Protein-protein interaction.