زیست فناوری دانشگاه تربیت مدرس
دوره 7، شماره 3، ویژه‌نامه 1395

PSNetAl: الگوریتمی جدید جهت ردیف‌بندی ساختاری چندتایی در پروتئین‌ها بر مبنای منطبق‌سازی گراف‌ها

نیلوفر نیک‌نام[1]، سید شهریار عرب[2*]، حسین نادری‌منش[3*]

1- دانشجوی دکتری بیوفیزیک، گروه بیوفیزیک، دانشکده علوم زیستی، دانشگاه تربیت مدرس، تهران
2- استادیار گروه بیوفیزیک، دانشکده علوم زیستی، دانشگاه تربیت مدرس، تهران
3- استاد گروه بیوفیزیک، دانشکده علوم زیستی، دانشگاه تربیت مدرس، تهران

* تهران، صندوق پستی 175-14115
* sh.arab@modares.ac.ir , naderman@modares.ac.ir
(دریافت مقاله: 94/8/17 پذیرش مقاله: 94/10/20)

چکیده- با افزایش روزافزون توالی‌ها و ساختارهای جدید پروتئینی به پایگاه‌های اطلاعاتی نظیر PDB، اهمیت مقایسه‌ی ساختاری پروتئین‌ها به منظور بررسی روابط تکاملی بین خانواده‌های مختلف، پیشگویی عملکرد در پروتئین‌های شرح نویسی شده و نیز طبقه‌بندی ساختار و فولد در پروتئین‌ها ضروری به نظر می‌رسد. برنامه‌های هم‌ردیفی ساختاری در پروتئین‌ها به دلیل حجم بالای محاسبات در مقایسه با روش‌های متداول ردیف‌بندی توالی‌ها کندتر هستند و جواب‌های تخمینی را ارائه می‌کنند. از این‌رو، طراحی الگوریتم‌های جدید در این زمینه یک مسئله‌ی باز محسوب می‌شود.

در این تحقیق، ارائه‌ی الگوریتمی جدید بر مبنای جورسازی گراف‌ها جهت انجام ردیف‌بندی‌های ساختاری چندتایی در پروتئین‌ها معرفی می‌شود. ورودی برنامه‌ی PSNetAl فایل‌های ساختاری پروتئین به فرمت PDB است. برای تمامی فایل‌های ساختاری پروتئین، گراف‌های غیرجهت‌دار و مبتنی بر فاصله ساخته می‌شود و با استفاده از الگوریتمی پیشرونده، هم‌ردیفی چندتایی شبکه‌ها انجام می‌شود. بزرگترین زیرگراف مشترک حاصل از هم‌ردیفی‌های چندتایی در انتهای برنامه شامل رأس‌های مشترک میان تمامی ساختارهای ورودی است. چنانچه میان ساختارهای ورودی شباهت‌های ساختاری و تکاملی وجود داشته باشد، انتظار می‌رود که موتیف-های ساختاری مشترک با انجام هم‌ردیفی چندتایی شبکه‌ها در بزرگترین زیرگراف مشترک قابل مشاهده باشند.

به منظور بررسی عملکرد برنامه، مجموعه داده‌ای شامل 76 خانواده‌ی پروتئینی با متوسط یکسانی بیش از 50% و حداقل دارای سه عضو از پایگاه داده‌ی HOMSTRAD استخراج شد. نتایج حاصل از این خانواده‌ها نشان می-دهد که هم‌ردیفی‌های ساختاری به درستی انجام شده است و در نتیجه‌ی آن در 67 خانواده از مجموع 76 خانواده بیش از 90% موتیف‌های ساختاری در بزرگترین زیرگراف مشترک دیده می‌شود.

کلید واژگان: هم‌ردیفی ساختاری، شبکه‌های ساختاری پروتئین، موتیف‌های ساختاری، بزرگترین زیرگراف مشترک، گراف‌های غیرجهت‌دار و مبتنی بر فاصله.

1- مقدمه

عملکرد پروتئین به طور معنی‌داری تابعی از ساختار فضایی آن است. باور عمومی بر آن است که ساختار و فولد سه بعدی یک پروتئین تأثیر عمده‌ای بر توانایی آن در اتصال به پروتئین‌های دیگر، لیگاندها (داروها)، پایداری و به طور کلی سایر جنبه‌های عملکردی دیگر آن خواهد گذاشت. بنابراین بررسی شباهت‌های ساختاری میان پروتئین‌ها به عنوان ماشین‌های عملکردی سلول نقش مهمی در فهم فرایند حیات در یک سلول دارد [1].

یکی از روش‌های متداول بررسی و مقایسه‌ی شباهت بین ساختارهای پروتئینی، ردیف‌بندی است. ردیف‌بندی می‌تواند بر اساس توالی و یا ساختار به صورت دوتایی و یا چندتایی انجام شود. عمده‌ترین هدف ردیف‌بندی، شناسایی رزیدوهای همولوگ است. بر خلاف ردیف‌بندی توالی‌ها که محور آن ردیابی تغییرات تکاملی پروتئین از طریق بررسی روند تغییرات بیوشیمیایی اسیدهای آمینه است، ردیف‌بندی ساختاری امکان مقایسه‌ی ویژگی‌های توپولوژیک و فضایی رزیدوها را فراهم می‌سازد. در هم‌ردیفی ساختاری در صورتی دو زیرواحد، معادل در نظر گرفته می‌شوند که در بهترین حالت برهم‌نهی[1] دو ساختار، فاصله‌ی اقلیدوسی دو رزیدو در فضا از یکدیگر اندک و آرایش[2] و جهت‌گیری آن دو به مقدار قابل توجهی مشابه باشد. انتظار براین است هنگامی که توالی دو پروتئین شباهت زیادی با یکدیگر دارد، ساختار آن دو نیز به همان نسبت با یکدیگر شباهت داشته باشد. با توجه به اینکه در طول تکامل، ساختار پروتئین‌ها نسبت به توالی آن‌ها بیشتر حفظ شده است، در موارد متعدد دیده شده است که با وجود شباهت قابل توجه ساختارها، توالی‌های دو پروتئین شباهت اندکی دارند. بنابراین در مواردی‌که یکسانی اندکی بین توالی‌ها وجود دارد ساختار نسبت به توالی، اطلاعات بیشتری را در زمینه‌ی خصوصیات وعملکرد پروتئین ارائه می‌دهد [2-4]. پروتئین‌هایی با عملکرد مشابه که توالی آنها یکسانی اندکی را نشان می‌دهد، غالباً دارای یک هسته‌ی مشترک بوده که شامل رزیدوهایی است که برای حفظ انسجام و یکپارچگی فولد پروتئین ضروری است [5]. این قطعات ساختاری محافظت شده اصطلاحاً موتیف‌های ساختاری نامیده می‌شود و می‌توان از آنها به عنوان حداقل ملزومات ساختاری برای نسبت دادن یک عضو جدید به یک خانواده و یا ابرخانواده نام برد.

برای انجام هم‌ردیفی ساختاری چندتایی در پروتئین‌ها تاکنون چندین نرم‌افزار و برنامه‌ی تحت شبکه بر اساس الگوریتم‌های مختلف طراحی شده است که در آنها عموماً هندسه‌ی[3] اتم‌های CA در زنجیره‌ی اصلی[4] مبنای مقایسه قرار می‌گیرد [6] و از مدل‌های مختلف نمایش، توابع امتیازدهی متنوع و الگوریتم‌های بهینه‌سازی متفاوت استفاده می‌شود. CE [7] و DALI [8] دو روش متداول برای جستجوی شباهت در پایگاه‌های داده‌ی ساختاری هستند که برای مقایسه‌های دوتایی پروتئین‌ها طراحی شده‌اند. در هر دو روش در دو ساختار، قطعات متناظر که دارای فواصل درون مولکولی مشابه از اتم‌های کربن آلفا هستند، جستجو می‌شوند و سپس با استفاده از استراتژی‌های مختلف این قطعات به یکدیگر متّصل می‌شوند. روش‌هایی همانند FATCAT [9, 10] قادر هستند که زیردومین‌ها را در آرایش‌های نسبی مختلف که نتیجه انعطاف‌پذیری پروتئین و یا واگرایی تکاملی است، هم‌ردیف کنند. استراتژی دیگر این است که علاوه بر هندسه‌ی زنجیره‌ی اصلی، محیط فیزیکوشیمیایی هر رزیدو نیز در هم‌ردیفی دو ساختار دو ساختار در نظر گرفته شود. از این شیوه در روش SHEBA [11] استفاده می‌شود. در بعضی از ابزارها مانند MATRAS [12] در

[1] Superposition
[2] Orientation
[3] Geometry
[4] Back bone

مرحله‌ی اول هم‌ردیفی برای رسیدن به یک هم‌ردیفی بهینه از انطباق عناصر ساختاری دوم استفاده می‌شود. در مرحله‌ی بعد از خصوصیات محیطی و نیز فواصل اتم‌های کربن آلفا برای رسیدن به هم‌ردیفی نهایی استفاده می‌شود. هم‌ردیفی‌های ساختاری چندتایی بطور قابل ملاحظه‌ای دارای اطلاعات بیشتری از هم‌ردیفی‌های دوتایی هستند. هدف مشترک تمام روش‌های هم‌ردیفی چندتایی شناسایی مجموعه‌ای از رزیدوها (ستون‌هایی) در هر ردیف است که از نظر ساختاری شبیه هستند و بلوک‌های هم‌ردیف شده نشان دهنده‌ی قطعات مشابه محلی هستند [13]. برای هم‌ردیفی ساختاری پروتئین‌ها برنامه‌های مختلفی وجود دارد که تفاوت عمده‌ی آنها با یکدیگر در نحوه‌ی انجام هم‌ردیفی‌های دوتایی و چگونگی ادغام آنها با یکدیگر برای ساخت یک هم‌ردیفی چندتایی است [14-17].

سرعت اجرای الگوریتم‌های ردیف‌بندی ساختاری در مقایسه با الگوریتم‌های معمول ردیف‌بندی توالی کمتر است و راه حل دقیقی را نیز ارائه نمی‌دهد که ناشی از این واقعیت است که مقایسه‌ی ساختاری پروتئین‌ها در بیوانفورماتیک یک مسئله‌ی NP سخت محسوب می‌شود. از این رو در الگوریتم‌های ارائه شده برای کاهش هزینه‌ی محاسبات[5] عموماً از تلفیقی از روش‌های مکاشفه‌ای[6]، حریصانه[7] و برخی پارامترهای تجربی استفاده می‌شود. در حال حاضر توسعه و گسترش روش‌های محاسباتی برای انجام هم‌ردیفی ساختاری چندتایی به خصوص در مجموعه پروتئین‌هایی که از لحاظ تکاملی دور از هم هستند به عنوان یک چالش و مسئله‌ی باز محسوب می‌شود. از ردیف‌بندی ساختاری چندتایی می‌توان در طبقه‌بندی ساختاری پروتئین‌ها، فیلوژنی ساختاری، شناسایی جایگاه فعال آنزیم‌ها و هسته‌ی مشترک میان دسته‌های پروتئینی، شناسایی هومولوژی‌های دوربین ساختارها و پیش‌گویی ساختار/ عملکرد و درک بهتر محافظت‌شدگی و واگرایی تکاملی در پروتئین‌ها استفاده کرد.

در این مقاله الگوریتم PSNetAl معرفی می‌شود که با کاهش ساختار سه بعدی پروتئین به صورت یک نمایش دو بعدی به شکل گراف، جورسازی گراف‌ها و نیز تلفیق اطلاعات توپولوژی و شباهت‌های زیستی رأس‌ها (CA از هر اسید آمینه) ردیف‌بندی ساختاری چندتایی شبکه‌های پروتئینی را به صورت سرتاسری انجام می‌دهد. یکی از مهمترین کاربردهای این برنامه استخراج موتیف‌های ساختاری محافظت شده در یک خانواده‌ی پروتئینی است.

2- روش‌ها

PSNetAl الگوریتمی پیش‌رونده[8] برای هم‌ردیفی ساختاری چندتایی پروتئین‌ها است. این الگوریتم از نظر ساختار کلی وفادار به الگوریتم NetAl [18] است که الگوریتمی حریصانه برای جورسازی دوتایی گراف‌های حاصل از شبکه‌های برهم‌کنش پروتئین-پروتئین است. الگوریتم NetAl دارای ویژگی‌های مهمی است که باعث شد از آن برای ایجاد و گسترش هم‌ردیفی ساختاری چندتایی پروتئین استفاده شود. مهمترین مزیت این الگوریتم پیچیدگی زمانی کمترآن ($O(n^2 \log n + m^2 + nm\log n)$) نسبت به سایر الگوریتم‌های مشابه است [18]. دستیابی به بهترین جواب[9] ردیف‌بندی و همچنین یافتن بزرگترین زیرگراف مشترک و پیوسته در شبکه از ویژگی‌های دیگر الگوریتم NetAl است. ورودی‌های برنامه‌ی PSNetAl عبارتند از فایل‌های ساختاری پروتئین‌های مورد نظر در فرمت PDB (که بیش از دو ساختار را شامل می‌شود). قابل ذکر است که در تعداد پروتئین‌های ورودی به برنامه محدودیتی وجود ندارد.

همانطور که در شکل 1 دیده می شود در الگوریتم PSNetAL سه مرحله ی کلی دیده می شود:

[5] Computational cost
[6] Heuristic
[7] Greedy

[8] Progressive
9 Optimum local

2-1- ساخت شبکه‌ی غیرجهت‌دار

در این مرحله از فایل‌های ورودی به فرمت PDB، زنجیره ومدل‌های ساختاری مورد نظر استخراج شده و گراف‌هایی غیرجهت‌دار و مبتنی بر فاصله ساخته می‌شود. در این گراف‌ها اتم کربن آلفا از هر اسید آمینه، یک رأس از شبکه‌ی پروتئینی را تشکیل می‌دهد. برای ساخت شبکه، یک ماتریس فاصله‌ی n×n (n تعداد اسیدهای آمینه در هر پروتئین است) از فاصله‌ی تمامی جفت اتم‌های کربن آلفا نسبت به یکدیگر ایجاد می‌شود و سپس با استفاده از یک حد آستانه‌ی فاصله، ماتریس مجاورت از ماتریس فاصله ساخته می‌شود. از ماتریس مجاورت برای ساختن گراف استفاده می‌شود. در اینجا از حد آستانه‌ی 8 آنگستروم برای ساختن شبکه‌های پروتئینی [19] استفاده شد.

2-2- ساخت ماتریس شباهت زیستی

برای انجام عمل هم‌ردیفی، علاوه بر شبکه‌های پروتئینی به فایل‌های شباهت نیز نیاز است. فایل‌های شباهت از فایل‌های PDB ورودی بدست می‌آید. در ماتریس شباهت زیستی، بر حسب اینکه تشابه بین رأس‌ها بر چه اساسی تعریف شود، شباهت هر رأس از شبکه‌ی اول با تمام رأس‌های شبکه‌ی دوم محاسبه می‌شود. در ساده‌ترین حالت ممکن این شباهت می‌تواند بر حسب شباهت اسیدهای آمینه به یکدیگر در ماتریس‌های جابه‌جایی Pam و یا Blosum تعریف شود. در حالت‌های پیچیده‌تر می‌توان ترکیبی از چند خصوصیت ساختاری همانند سطح در دسترس، ساختار دوم و ... را در نظر گرفت. در این بررسی، ماتریس شباهت زیستی بر اساس شباهت اسیدهای آمینه در ماتریس Blosum62 تعریف شده است.

2-3- مرحله هم‌ردیفی چندتایی

هم‌ردیفی چندتایی از چند مرحله‌ی مستقل از هم تشکیل شده است که عبارت است از:

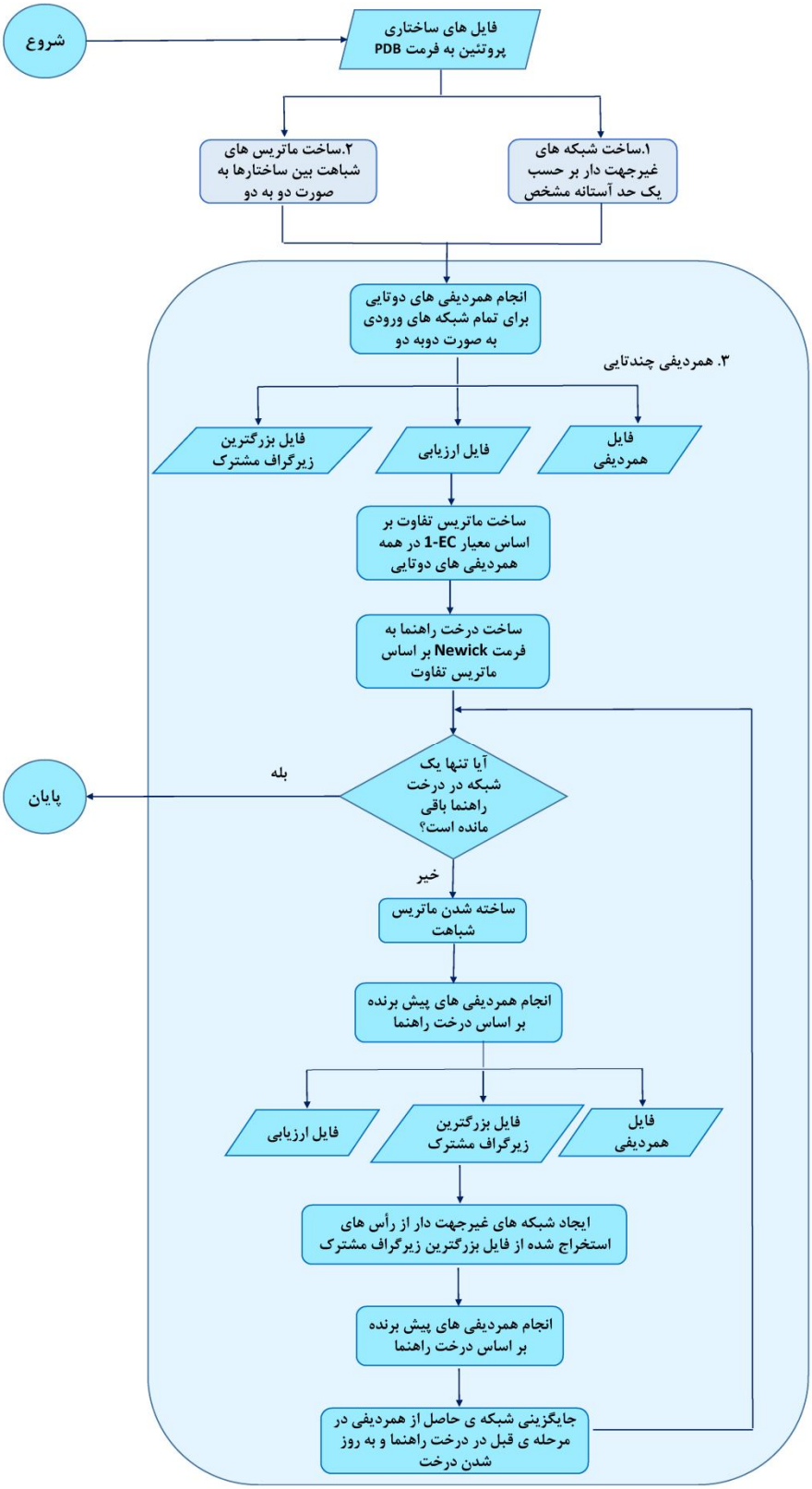
الف) مرحله‌ی هم‌ردیفی‌های دوتایی: برنامه‌ی PSNetAl برای انجام هم‌ردیفی‌های دوتایی از برنامه NetAl استفاده می‌کند. فایل‌های ورودی به برنامه عبارتند از فایل‌های شبکه دو پروتئین و فایل‌های ماتریس شباهت. پارامترهای ورودی به برنامه که عبارتند از پارامترهای a و c. پارامتر a که وزن ماتریس شباهت توپولوژی در مقابل شباهت زیستی را مشخص می‌کند. برای بدست آوردن شباهت توپولوژی پارامتر c وزن شباهت بین دو رأس و یا همسایه‌های آن دو را مشخص می‌کند.

در انجام هر هم‌ردیفی دوتایی، ابتدا رأس‌های مشابه در ساختار گراف شناسایی شده و سپس با یکدیگر منطبق می‌شود. در گام‌های بعد با افزودن همسایه‌های نزدیک، منطبق‌سازی گسترش می‌یابد. برای یافتن رأس‌های مشابه با یکدیگر در هر شبکه، به ناچار شباهت رأس‌ها با یکدیگر باید کمّی شوند. مقدار شباهت برای دو رأس i و j از دو شبکه n و m در ماتریس شباهت (S) طبق فرمول (1) محاسبه می‌شود:

$$S(i,j) = aT(i,j) + (1-a)B(i,j) \qquad (1)$$

ماتریس شباهت در برنامه‌ی PSNetAl از مجموع دو ماتریس شباهت توپولوژیکی (T) و شباهت زیستی (B) با ضرایب a و (1-a) تشکیل شده است. پارامتر a در این فرمول عددی بین صفر و یک است که وزن شباهت زیستی و توپولوژیک را در هم‌ردیفی‌ها مشخص می‌کند.

ماتریس امتیازدهی ردیف‌بندی (A) نیز یک ماتریس n×m است که در آن A_{ij} (i رأسی از شبکه‌ی n و j رأسی از شبکه‌ی m است) امتیاز نهایی هم‌ردیفی برای دو رأس i و j از دو شبکه‌ی n و m را نشان می‌دهد. این ماتریس از مجموع ماتریس شباهت (S) و ماتریس برهم‌کنش‌ها (I) ساخته می‌شود. فرایند انتخاب بهترین کاندیدا برای هم‌ردیفی سرتاسری از بین جفت رأس‌های موجود بین دوشبکه از طریق این ماتریس انجام می‌شود.



شکل 1 مراحل کلی الگوریتم PSNetAl

پارامتر c میزان وزن رأس‌های هم‌ردیف شده را در تقابل با میزان وزن رأس‌های همسایه‌های آنها با یکدیگر در هم‌ردیفی مشخص می‌کند. اگر c=0 باشد به این معنا است که برنامه فقط شباهت دو رأس مورد مقایسه را در نظر می‌گیرد و اگر c=1 باشد، برنامه فقط شباهت همسایه‌های رأس‌های اصلی را در نظر می‌گیرد.

در این بررسی برای یکسان شدن شرایط مورد نظر برای تمام خانواده‌های پروتئینی، تمام هم‌ردیفی‌های دوتایی و چندتایی با پارامتر a = 0/3 و c = 0/7 انجام شد. پس از انجام هر هم‌ردیفی، نتیجه به صورت چند فایل خروجی مجزا نشان داده می‌شود. این فایل‌ها عبارتنداز:

فایل هم‌ردیفی (alignment.): که در آن تمام رأس‌های جفت شده با یکدیگر از دو شبکه به همراه امتیاز هم‌ردیفی آنها نشان داده می‌شود.

فایل ارزیابی (eval.): که در آن صحّت انجام هم‌ردیفی از طریق کمیّت درستی یال‌ها[1] (EC) و به صورت درصد بیان بیان می‌شود. درستی یال، درصد یال‌هایی از شبکه‌ی اول را بیان می‌کند که با یال‌هایی از شبکه‌ی دوم هم‌ردیف شده‌اند [20]. EC طبق فرمول (2) تعریف می‌شود [21]:

$$C = \frac{|\{u,v\} \in E_1 : (g(u), g(v)) \in E_2}{|E|} \times 100\% E \quad (2)$$

بدیهی است که هر چقدر درصد درستی یال‌ها بزرگتر باشد، هم‌ردیفی با صحّت بیشتری انجام شده است و شبکه‌ی G_1 و G_2 از نظر توپولوژیک شبیه‌ترند. همچنین در این فایل تعداد رأس‌ها و یال‌های بزرگترین زیر گراف مشترک حاصل از هم‌ردیفی نیز ثبت خواهد شد. اندازه‌ی بزرگترین زیرگراف مشترک حاصل از هم‌ردیفی دو شبکه کوچکتر و یا مساوی با اندازه‌ی گراف کوچکتر خواهد بود و هر چقدر اندازه‌ی بزرگترین زیرگراف مشترک به اندازه‌ی شبکه‌های ورودی اولیه به برنامه نزدیک‌تر باشد نشان دهنده‌ی این است که دو شبکه شباهت بیشتری با یکدیگر دارند.

فایل بزرگترین زیرگراف مشترک[2]: این فایل نشان می‌دهد که بزرگترین زیرگراف مشترک حاصل از هم‌ردیفی دارای چه رأس‌هایی است و نیز رأس‌های بزرگترین زیرگراف مشترک از شبکه‌ی کوچکتر با کدام یک از رأس‌های شبکه‌ی بزرگتر هم‌ردیف شده‌اند. امتیاز هم‌ردیفی دو به دوی رأس‌ها از دو شبکه‌ی مختلف در این فایل نمایش داده می‌شود. در شکل 2 نحوه‌ی تشکیل بزرگترین زیرگراف مشترک به صورت شماتیک نمایش داده شده است.

در این مرحله هم‌ردیفی برای تمامی شبکه‌های ورودی به برنامه به صورت دو به دو انجام می‌شود.

ب) ساخت ماتریس تفاوت: برای بدست آوردن درخت راهنما، می‌بایست ماتریس تفاوت که دربر گیرنده‌ی میزان اختلاف جفت ساختارها از یکدیگر است، محاسبه شود. بنابراین برای بدست آوردن توپولوژی ساختارها در درخت راهنما، از مقدار 1-EC که نشان دهنده‌ی اختلاف ساختارها از یکدیگر است استفاده می‌شود.

ت) ساخت درخت راهنما بر اساس درصد صحّت هم‌ردیفی یال‌ها (EC): برای ساخت درخت راهنما از الگوریتم‌های متفاوتی می‌توان استفاده نمود. در این تحقیق از الگوریتم UPGMA استفاده شد. درخت راهنما جهت معرفی به برنامه، به فرمت Newick (به صورت رشته[3]) ذخیره می‌شود. لازم به ذکر است که توپولوژی درخت به هیچ عنوان دربردارنده‌ی اطلاعات فیلوژنی نبوده و از آن صرفاً به عنوان یک راهنما برای ترتیب هم‌ردیفی‌های دوتایی به شکل پیشرونده استفاده می‌شود.

ث) بر اساس درخت راهنما در هر نوبت شبیه‌ترین شبکه‌ها با یکدیگر هم‌ردیف می‌شوند.

ج) فایل بزرگترین زیرگراف مشترک شامل رأس‌های مشترک بین دو شبکه است که برهم منطبق شده‌اند.

[1] Edge correctness

[2] Largest common subgraph

[3] String

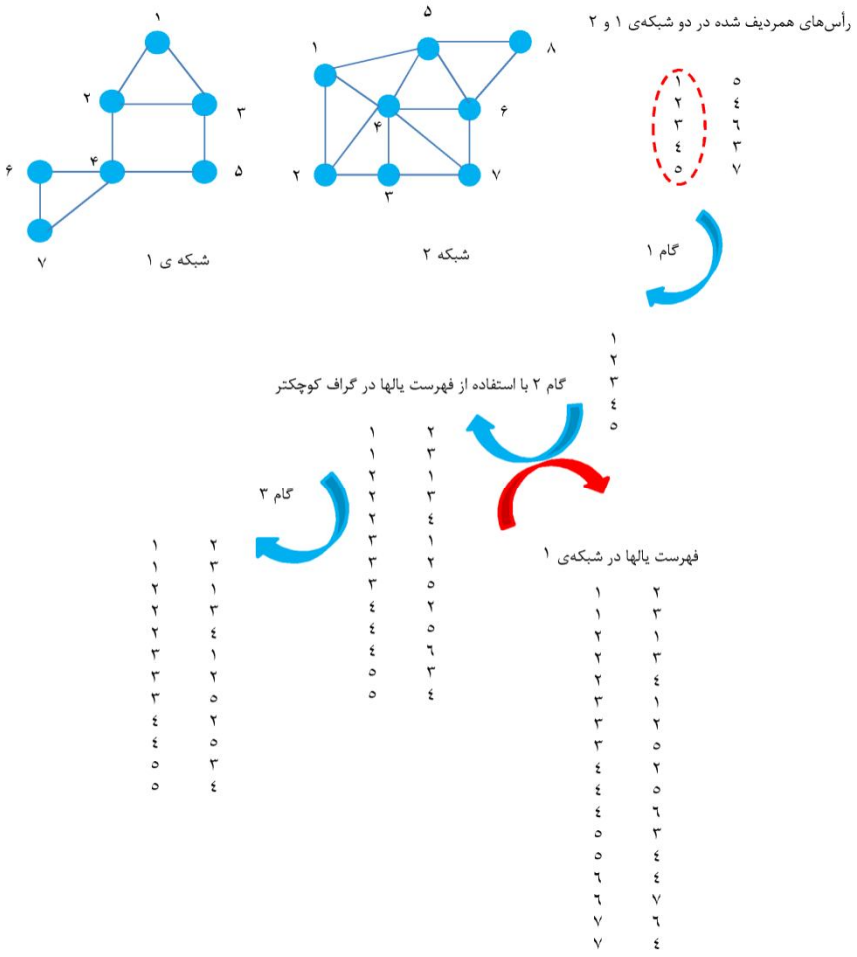
از این رأس‌ها می‌توان برای استخراج شبکه‌ی جدید از دو شبکه‌ی والد استفاده کرد. شبکه‌ی بدست آمده در درخت راهنما جایگزین دو شبکه‌ی والد خود شده و بدین ترتیب پس از هر دور هم‌ردیفی، درخت روزآمد می‌شود و توپولوژی آن تغییر می‌کند. تا زمانی‌که بیش از یک شبکه در توپولوژی درخت وجود داشته باشد، هم‌ردیفی ادامه می‌یابد. هنگامی‌که تنها یک شبکه در درخت راهنما وجود داشته باشد، برنامه‌ی هم‌ردیفی چندتایی پایان می‌یابد.

به این ترتیب، آنچه در پایان هم‌ردیفی‌های چندتایی باقی می‌ماند، مجموعه‌ای از رأس‌های مشترک بین پروتئین‌های مختلف است که دربرگیرنده‌ی شباهت‌های توپولوژیک و زیستی بین آنهاست. از آنجا که موتیف‌های ساختاری دارای ویژگی‌های توپولوژیک و فضایی محافظت شده میان اعضای مختلف خانواده‌ها هستند، انتظار می‌رود که در صورت درستی هم‌ردیفی چندتایی در شبکه‌های ورودی، این موتیف ها در میان رأس‌های مشترک موجود در آخرین هم‌ردیفی دیده شوند. در بخش نتایج بیشتر در این مورد توضیح داده خواهد شد.

3- تهیه‌ی مجموعه‌ی داده

هم‌ردیفی ساختاری پروتئین‌ها را می‌توان به نوعی یک هم‌ردیفی توالی در نظر گرفت که در آن انتقال هندسی نیز به صورت همزمان به یافتن و هم‌ردیف کردن اتم‌های کربن آلفا از اسیدهای آمینه‌ی معادل کمک می‌کند [22]. برای تهیه‌ی یک مجموعه‌ی مرجع برای بررسی و ارزیابی درستی و دقّت برنامه‌های هم‌ردیفی ساختاری در پروتئین‌ها باید از پایگاه‌های داده حاوی هم‌ردیفی‌های چندتایی توالی‌ها استفاده شود. مزیت مهم این پایگاه‌های داده آن است که با استفاده از اصل محافظت شدگی بیشتر ساختار نسبت به توالی در طول تکامل با استفاده از ویژگی‌های مختلف ساختاری و نیز اعمال نظر مستقیم افراد خبره هم‌ردیفی‌های چندتایی بر مبنای ویژگی‌های ساختاری ایجاد شده‌است. با کمک گرفتن از اطلاعات موجود در این پایگاه‌های داده‌ی مرجع برنامه‌های مختلف هم‌ردیفی ساختاری ارزیابی می‌شود [16،22-24].

در پایگاه داده‌ی MegaMotifBase با استفاده از هم‌ردیفی‌های ساختاری چندتایی خانواده‌های مختلف که از پایگاه HOMSTRAD جمع‌آوری شده است و با استفاده از ویژگی‌های ساختاری مختلف، موتیف‌های ساختاری مشترک میان اعضای یک خانواده تعیین می‌شود. موتیف‌های ساختاری علاوه بر اینکه در میان تمامی اعضای یک خانواده دارای ویژگی‌های ساختاری و توپولوژیک مشترک می‌باشند، مناطق ساختاری محافظت شده‌ی هر خانواده را نیز تشکیل می‌دهند و در هم‌ردیفی‌های ساختاری نیز با تشکیل ستون‌هایی از رزیدوهای یکسان در میان تمامی اعضا، بلاک‌های ساختاری را در هر هم‌ردیفی چندتایی شکل می‌دهند. از اطلاعات موجود در مورد موتیف‌های ساختاری مشترک در میان اعضای هر خانواده در پایگاه داده MegaMotifBase و نیز هم‌ردیفی‌های موجود در پایگاه داده‌ی HOMSTRAD برای بررسی درستی هم‌ردیفی‌های ساختاری انجام شده توسط برنامه‌ی PSNetAl می‌توان استفاده کرد. برای بررسی درستی این برنامه می‌توان فرض کرد که در صورت انجام هم‌ردیفی چندتایی درست در هر خانواده، می‌بایست رزیدوهای تشکیل دهنده‌ی مناطق ساختاری مشترک در بزرگترین زیرگراف حاصل از آخرین هم‌ردیفی دوتایی نیز دیده شود. از این رو، برای بررسی درستی برنامه می‌توان تعداد ستون‌های محافظت شده در هم‌ردیفی حاصل از برنامه‌ی PSNetAl را با هم‌ردیفی‌های مرجع موجود در پایگاه‌های HOMSTRAD و یا MegaMotifBase مقایسه کرد [25]. هر چه تعداد ستون‌های محافظت شده در هم‌ردیفی چندتایی حاصل به تعداد ستون‌های رزیدوهای محافظت شده در هم‌ردیفی ساختاری مرجع نزدیک‌تر باشد (به عبارتی تمام موتیف‌های ساختاری در بزرگترین زیرگراف مشترک وجود داشته باشد) می‌توان ادّعا کرد که هم‌ردیفی درست‌تری انجام شده است.



شکل 2 نحوه‌ی تشکیل یک گراف جدید از فایل بزرگترین زیرگراف مشترک

HOMSTRAD، شامل مجموعه‌ای ازهم‌ردیفی‌های چندتایی توالی‌های پروتئینی است که در انجام هم‌ردیفی علاوه بر لحاظ کردن محافظت‌شدگی توالی‌ها، ویژگی‌های ساختاری همانند سطح در دسترس، محتوای ساختار دوم، الگوی پیوندهای هیدروژنی، پیوندهای دی‌سولفیدی و ... در نظر گرفته شده است. از داده‌های موجود در این پایگاه داده به عنوان یک استاندارد و هم‌ردیفی مرجع برای بررسی هم‌ردیفی‌های ساختاری استفاده می‌شود. در HOMSTRAD، هم‌ردیفی‌های موجود در سطح خانواده‌های پروتئینی است و در آن ساختارهای پروتئینی موجود از بانک اطلاعاتی پروتئین PDB[1] جمع‌آوری و در بیش از 800 خانواده‌ی چند عضوی طبقه‌بندی شده است. از اطلاعات موجود در این پایگاه داده می‌توان برای مدل‌سازی مقایسه‌ای و شناسایی پروتئین‌هایی با رابطه‌ی تکاملی دور استفاده کرد [26].

به این منظور از میان 266 خانواده‌ی پروتئینی با بیش از 50 درصد یکسانی در توالی، 76 خانواده با بیش از 2 عضو برای انجام هم‌ردیفی‌های چندتایی انتخاب شد. همان‌طور که نمودار 1 نشان می‌دهد تعداد خانواده‌های پروتئینی انتخاب شده از نظر درصد یکسانی توالی بین

[1] Protein Data Bank

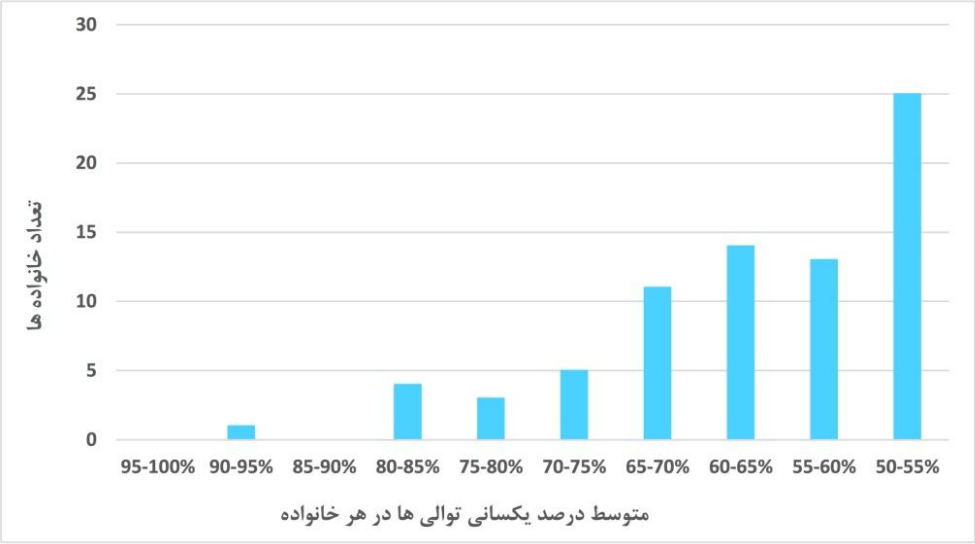
ساختارهای تشکیل دهنده دارای پراکنش یکسان نیست. خانواده‌هایی با درصد یکسانی توالی در بازه 95-90 درصد و نیز 90-85 درصد (به ترتیب با یک و صفر عضو) دارای کمترین تعداد وخانواده‌هایی با درصد یکسانی توالی در محدوده‌ی 55-50 درصد (با 24 عضو) دارای بیشترین تعداد در مجموعه خانواده‌های مورد بررسی می‌باشد.

4- نتایج

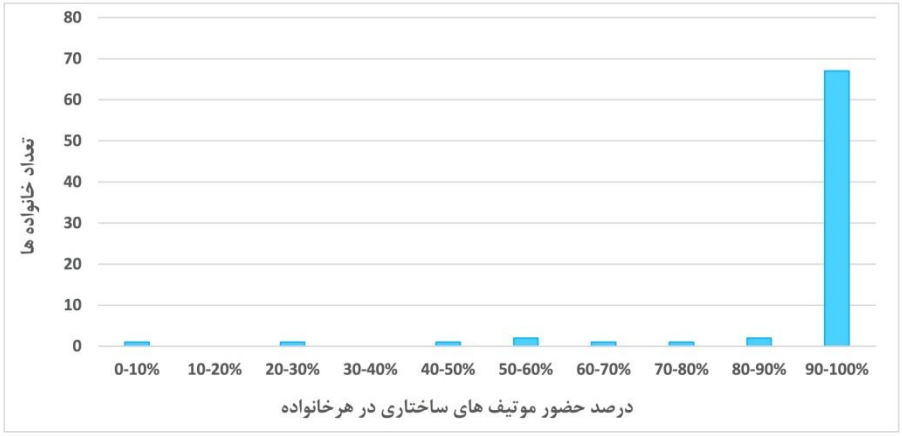
برنامه‌ی هم‌ردیفی چندتایی روی 76 خانواده‌ی پروتئینی با درصد یکسانی توالی بیش از 50 درصد و تعداد بیش از دو عضو اجرا شد که نتایج آن در جدول 1 نشان داده شده است. طبق این نتایج و همان‌طور که در نمودار 2 دیده می‌شود در 67 مورد (88 درصد) از مجموع 76 خانواده‌ی پروتئینی بررسی شده درصد فراوانی موتیف‌های ساختاری در بزرگترین زیرگراف مشترک حاصل از هم‌ردیفی بیش از 90% است. حضور 100% موتیف‌های ساختاری در بزرگترین زیرگراف مشترک حاصل از هم‌ردیفی به این معنا است که برنامه توانسته است رأس‌های پروتئین‌های موجود در هر خانواده را به بهترین شکل بر هم منطبق کند. در 2 خانواده (2/63 درصد)، بیش از 80 درصد موتیف‌های ساختاری در بزرگترین زیرگراف مشترک حضور داشته است. در 7 خانواده (9 درصد) کمتر از 70 درصد از موتیف‌های ساختاری در بزرگترین زیرگراف مشترک دیده شده است.

در بسیاری از خانواده‌ها با درصد یکسانی کم توالی‌ها، هم‌ردیفی‌های ساختاری با مقادیر EC زیاد انجام می‌شود. و در برخی از خانواده ها با درصد یکسانی توالی بالا گاهاً هم‌ردیفی ساختاری با مقادیر بالای EC انجام نمی‌شود (نمودار 3).

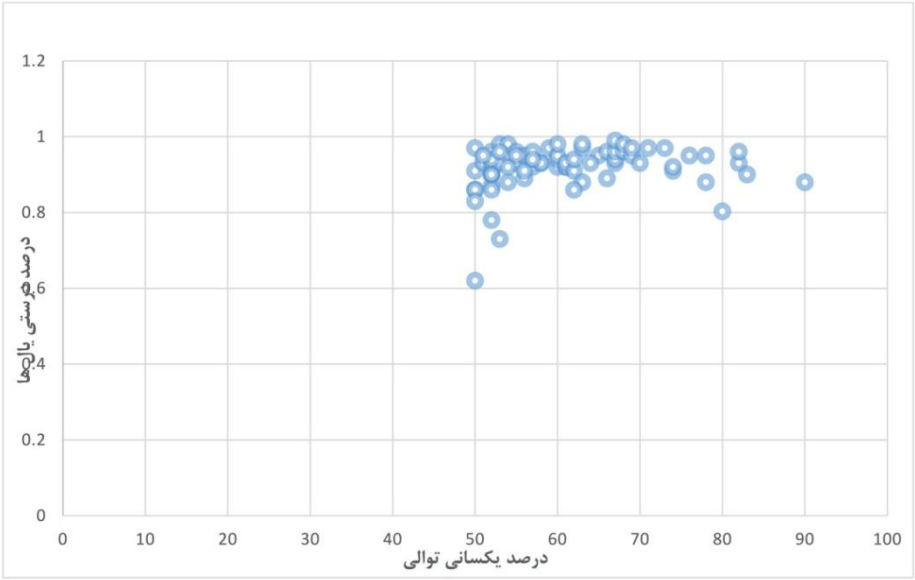
برای مثال در خانواده‌ی Paralytic/GBP/PSP peptide با 83 درصد یکسانی توالی، طول متوسط اعضا 23 است. با وجود درصد یکسانی بالایی که بین توالی‌های اعضای این خانواده وجود دارد، تنها 83 درصد از موتیف‌های ساختاری در زیرگراف مشترک دیده می‌شود. با توجه به کوچک بودن اعضای این خانواده نتیجه‌ی بدست آمده دور از انتظار نیست.



نمودار 1 این نمودار، تعداد خانواده‌های پروتئینی را که دارای درصد یکسانی توالی خاص هستند، نشان می‌دهد. خانواده‌هایی دارای درصد یکسانی توالی بین 55-50 گروهی هستند که بیشترین تعداد خانواده‌های مورد مطالعه را پوشش می‌دهد.



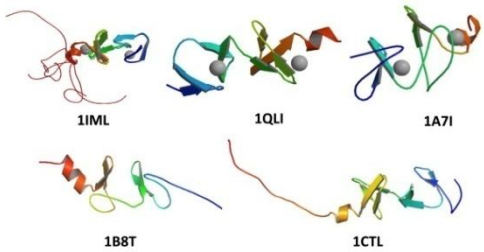
نمودار 2 درصد حضور موتیف‌های ساختاری دربزرگترین زیرگراف مشترک حاصل از هم‌ردیفی چندتایی در خانواده‌های پروتئینی



نمودار 3 رابطه‌ی بین درصد یکسانی توالی‌ها درخانواده‌های پروتئینی و درصد درستی یال‌ها در هر هم‌ردیفی چندتایی

هر چقدر طول رشته‌ی پروتئینی کوچکتر باشد، ساختار پروتئین دارای پیچش و فشردگی[1] کمتری خواهد بود. کاهش تماس‌های مؤثر هر رزیدو با رزیدوهای اطراف خود در ساختار، باعث کم شدن تعداد همسایه‌های آن رأس وکاهش خصوصیت توپولوژیک آن در گراف بدست آمده خواهد شد.

وجود ساختارهای متغیر (شکل‌های 3 و 4) در هر خانواده، تشخیص رأس‌های مشابه را در بین شبکه‌های ساختاری پروتئین دشوار خواهد کرد.



شکل 3 اعضای خانواده‌ی Zinc Binding domain present in Lin-11, Isl-1, Mec-3. در شکل، کوچک بودن و متغیر بودن ساختارها در میان اعضای این خانواده کاملاً مشخص است.

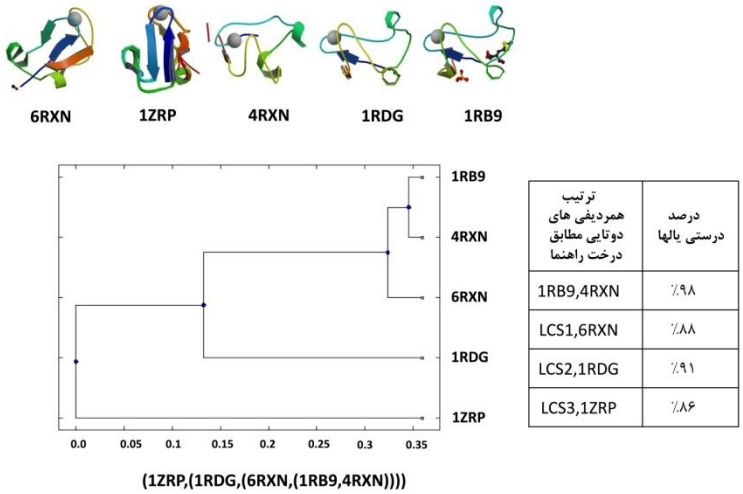
توأم شدن اندازه‌ی کوچک توالی‌های پروتئینی و وجود ساختارهای متغیر با شباهت اندک در یک خانواده با کم

[1] Fold & Packing

بودن درصد یکسانی توالی میان اعضا، تأثیر قابل توجهی بر کیفیت هم‌ردیفی رأس‌ها در دو گراف ساختاری خواهد داشت. کیفیت هم‌ردیفی در 3 خانواده‌ی Ruberdoxin با متوسط درصد یکسانی 62 در توالی‌های اعضا و خانواده‌ی Domains Containing Gla با متوسط یکسانی توالی 54 درصد و Zinc Binding domain present in Lin-11, Isl-1,Mec-3، با متوسط درصد یکسانی 53 در توالی‌ها به خوبی گویای این واقعیت است (نتایج هم‌ردیفی در جدول 1 قابل مشاهده است). در این خانواده‌ها با وجود درصدهای بالای درستی یال‌ها در هم‌ردیفی آخر، به دلیل فقدان ویژگی‌های توپولوژیک قوی در رأس‌های شبکه‌ی پروتئینی، تنها نیمی از موتیف‌های ساختاری در بزرگترین زیرگراف مشترک حاصل از هم‌ردیفی مشاهده می‌شود.

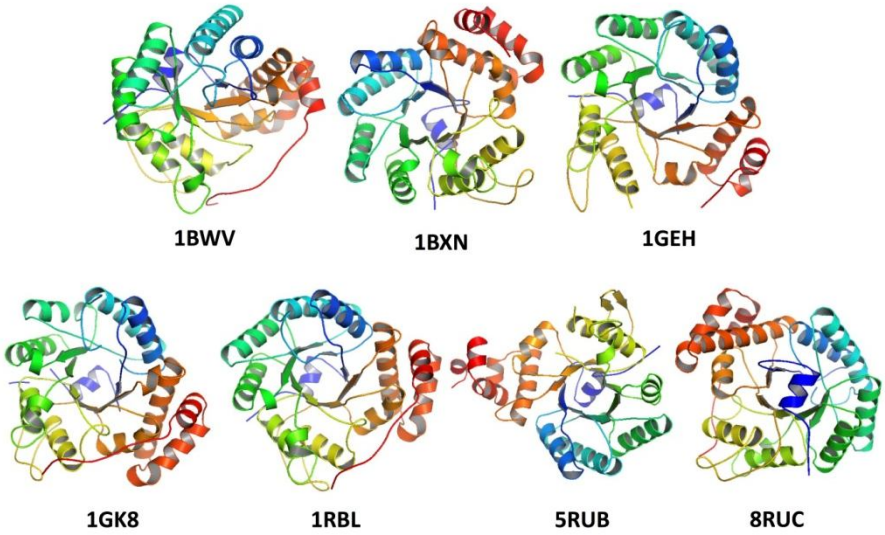
همان‌طور که انتظار می‌رود متوسط درصد حضور موتیف‌های ساختاری در خانواده‌هایی با درصد یکسانی توالی در بازه‌ی 55-50 نسبت به سایر خانواده‌ها کاهش می‌یابد. باید به این نکته توجه داشت که درصد یکسانی توالی‌ها در یک خانواده میانگینی از درصد یکسانی توالی‌ها میان اعضای آن خانواده است که بعضاً می‌تواند محدوده‌ی عددی متغیری را دربر گیرد. در خانواده‌ی ribulose bisphosphate carboxylase large chain این مورد به خوبی دیده می‌شود (شکل 5).

در این خانواده با وجود شباهت ساختاری اعضا با یکدیگر، محدوده‌ی تغییرات یکسانی توالی‌ها بین اعضا، 27-80 درصد است. متنوع بودن درصد یکسانی توالی‌ها بین اعضا باعث می‌شود که در هم‌ردیفی‌های آخر در درخت راهنما که میزان تفاوت بین دو ساختار بیشتر است درصد درستی یال‌ها کاهش یابد و متعاقب آن اندازه‌ی بزرگترین زیرگراف مشترک در هم‌ردیفی‌های متوالی کوچکتر شود. از این رو تنها 77% از موتیف‌های ساختاری در بزرگترین زیرگراف مشترک حاصل از هم‌ردیفی دیده می‌شود و 30 درصد بقیه در هم‌ردیفی‌های قبل حذف می‌شود. در اینجا با توجه به شباهت قابل توجه ساختارها به یکدیگر، برای دستیابی به یک هم‌ردیفی بهینه نیاز است که وزن شباهت‌های توپولوژیک در هم‌ردیفی‌ها افزایش داده شوند.



ترتیب هم‌ردیفی های دوتایی مطابق درخت راهنما	درصد درستی یالها
1RB9,4RXN	٪۹۸
LCS1,6RXN	٪۸۸
LCS2,1RDG	٪۹۱
LCS3,1ZRP	٪۸۶

شکل 4 (بالا) خانواده‌ی Ruberdoxin با درصد یکسانی 62 در میان اعضا. (پائین). درخت راهنما و جدول مقادیر درستی یال‌ها برای هم‌ردیفی‌های دوتایی پیشبرنده برای خانواده‌ی Ruberdoxin. با وجود مقادیر بالای درصد درستی یال‌ها برای هر هم‌ردیفی، به دلیل فقدان ساختار مشخص در نیمی از اعضای این خانواده، PSNetAl تنها توانسته است نیمی از موتیف‌های ساختاری موجود در این خانواده را در آخرین زیرگراف مشترک حاصل حفظ کند. منظور از LCS در ستون اول جدول، بزرگترین زیرگراف مشترک حاصل از هم‌ردیفی است.



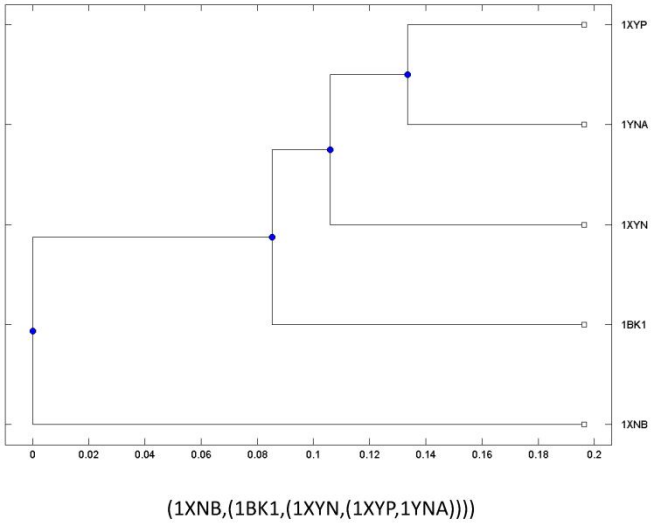
شکل 5 اعضای ribulose bisphosphate carboxylase large chain با درصد یکسانی 50 در توالی‌ها.

قابل ذکر است که در مواردی که یکسانی توالی اعضای خانواده‌ها کاهش می‌یابد (در اینجا در محدوده‌ی 50 درصد) با تغییر پارامترهای ورودی a و c می‌توان جواب‌های بدست آمده را بهبود بخشید. در این موارد با کم کردن مقدار پارامتر a که وزن ماتریس توپولوژی را مشخص می‌کند و افزایش مقدار پارامتر c که وزن شباهت همسایه‌های دو رأس را در ماتریس شباهت توپولوژیک تعیین می‌کند، می‌توان تأثیرات ناشی از کم بودن درصد یکسانی توالی در ساختارها را کاهش داد و به هم‌ردیفی‌های بهتری رسید. خانواده‌ی glycosyl hydrolase family 11 با یکسانی توالی 50 درصد در بین اعضا در این مورد مثال مناسبی است. همان‌طور که در شکل 6 دیده می‌شود در آخرین هم‌ردیفی پیشرونده طبق درخت راهنما به دلیل اختلاف ساختارها با یکدیگر و کاهش درصد یکسانی توالی‌ها، درصد درستی یال‌ها به صورت چشمگیری کاهش یافته و باعث کم شدن تعداد رأس‌های مشترک در زیرگراف ساخته شده می‌شود. نهایتاً درصد حضور موتیف ساختاری مشترک میان ساختارها درآخرین زیرگراف مشترک صفر خواهد بود. مطابق با ادعای بالا، در این خانواده با تغییر مقدار پارامترهای a و c (a = 0/2 و c =0/8) و افزایش وزن شباهت توپولوژیک در هم‌ردیفی آخر می‌توان بیش از 90 درصد از موتیف‌های ساختاری مشترک میان اعضا را در بزرگترین زیرگراف مشترک مشاهده کرد.

5- بحث در نتایج

هیچ تعریف آماری روشنی از عوامل ایجاد کننده‌ی شباهت‌های قابل ملاحظه در ساختارهای پروتئینی وجود ندارد که عمدتاً ناشی از سه دلیل عمده است. اول اینکه ساختار پروتئین بیش از آنکه تحت تأثیر توالی باشد به‌وسیله‌ی نیروهای فیزیکی و شیمیایی محدود[1] می‌شود. دوم آنکه هیچ تعریف دقیقی از یک هم‌ردیفی سه بعدی بهینه وجود ندارد. مورد سوم ناشی از این حقیقت است که مشخص کردن ساختارهای تصادفی یک پروتئین با دشواری همراه است. عدم قطعیت درباره‌ی دامنه‌ی ساختارهای پروتئینی محتمل و فقدان یک تعریف ریاضی از ساختارهای تصادفی، درک نظری اینکه چقدر احتمال دارد که دو پروتئین به صورت مستقل منجر به ساختارهای مشابه شوند را دشوار می‌کند.

[1] Constrain



درصد درستی یال‌ها	ترتیب هم‌ردیفی‌های دوتایی مطابق درخت راهنما
%93	1XYP,1YNA
%89	LCS1,1XYN
%90	LCS2,1BK1
%62	LCS3,1XNB

شکل 6 درخت راهنما در خانواده‌ی glycosyl hydrolase family 11. ترتیب هم‌ردیفی‌های دوتایی بین اعضا به صورت پیشرونده در جدول سمت راست تصویر مشاهده می‌شود. منظور از LCS بزرگترین زیرگراف مشترک بدست آمده از هم‌ردیفی در مرحله‌ی قبل است.

در حالی که این مسئله فرایند مقایسه‌ی ساختاری در پروتئین‌ها را چالش برانگیز می‌سازد، فقدان یک تعریف جامع از یک هم‌ردیفی بهینه مانع از آن می‌شود که بتوان به صورت تجربی توزیع امتیازهای شباهت ساختاری را تعیین و روش‌های مختلف را با یکدیگر مقایسه کرد [28،27] به همین دلیل برای ارزیابی برنامه‌های مختلف ناگزیر به مقایسه با یک حالت مرجع هست که در مورد آن توضیح داده شده است.

برنامه‌ی PSNetAl دارای دو کمیّت است که با استفاده از آنها می‌توان در مورد کیفیت هم‌ردیفی انجام شده اظهار نظر کرد. این دو فاکتور عبارتند از درصد درستی یال‌ها (EC) و نیز تعداد رأس‌های تشکیل دهنده‌ی موتیف‌های ساختاری در بزرگترین زیرگراف مشترک. با توجه به نتایج موجود در جدول 1 (فایل ضمیمه) می‌توان گفت که بین درصد درستی یال‌ها در هر خانواده با تعداد اعضای بزرگترین زیرگراف مشترک در همان خانواده و نیز تعداد موتیف‌های ساختاری موجود رابطه‌ی مستقیم وجود دارد. از آنجا که درصد درستی یال‌ها در هر هم‌ردیفی نشان دهنده‌ی شباهت توپولوژیک شبکه‌های هم‌ردیف شده با یکدیگر است می‌توان انتظار داشت که هر چقدر که معیار درستی یال‌ها به یک (100 درصد) نزدیک‌تر باشد دو شبکه به یکدیگر شبیه‌تر، نتیجه‌ی حاصل از هم‌ردیفی‌ها مطلوب‌تر، زیرگراف مشترک بزرگتر و نیز تعداد موتیف-های ساختاری بیشتری در زیرگراف نهایی قابل مشاهده است. بر اساس نتایج بدست آمده در رابطه با پروتئین‌های مختلف با مقادیر شباهت ساختاری متفاوت و نیز اندازه-های متنوع به صورت تجربی می‌توان گفت که هنگامی‌که مقدار درصد درستی یال‌ها بیش از 80 درصد باشد، می‌توان انتظار داشت که هم‌ردیفی با دقت مناسب انجام شده است. در مواردی که درصد درستی یال‌ها در هر هم‌ردیفی کمتر از 70 درصد باشد، غالباً جواب‌های بدست آمده صحیح نمی‌باشد. از سوی دیگر، با توجه به تعریف موتیف‌های ساختاری حاصل از هم‌ردیفی‌های ساختاری چندتایی در پروتئین‌ها می‌توان گفت که این رزیدوها علاوه بر مشابهت در توالی، دارای ویژگی‌های توپولوژیک مشترک و محافظت شده نسبت به سایر رزیدوها هستند؛ بنابراین انتظار می‌رود که این رزیدوها در صورت درستی هم‌ردیفی ساختاری در فایل بزرگترین

زیرگراف مشترک حاصل از هم‌ردیفی چندتایی باقی بمانند. در این بررسی، از دو کمیت درصد درستی یال‌ها و نیز تعداد موتیف‌های ساختاری مشاهده شده در بزرگترین زیرگراف مشترک نسبت به هم‌ردیفی‌های ساختاری مرجع موجود در پایگاه داده‌های HOMSTRAD و MegaMotifBase به عنوان دو معیار برای ارزیابی درستی هم‌ردیفی‌های صورت گرفته استفاده شد.

بین درصد یکسانی توالی‌ها در هر خانواده و نیز درصد درستی یال‌ها (EC%) ارتباط مستقیمی دیده نشد. واضح است که هر چقدر درصد یکسانی توالی‌ها بین اعضای یک خانواده بیشتر باشد با سهولت و اطمینان بیشتری می‌توان رأس‌های آن‌ها را بر یکدیگر منطبق کرد.

برای ارزیابی درستی هم‌ردیفی ساختاری چندتایی در برنامه‌ی PSNetal، این برنامه روی 76 خانواده‌ی پروتئینی با درصد یکسانی توالی بیش از 50 درصد اجرا شد. درصد حضور موتیف‌های ساختاری و یا به عبارتی درصد حضور ستون‌های محافظت شده از رزیدوهای کلیدی در هم‌ردیفی چندتایی که یکی از شاخص‌های نشان دهنده‌ی یک هم‌ردیفی صحیح است [29]، محاسبه شد. درصد حضور ستون‌های محافظت شده در هم‌ردیفی‌های چندتایی انجام شده توسط PSNetal در هر خانواده نسبت به هم‌ردیفی‌های مرجع محاسبه شد که نتیجه‌ی آن در جدول 1 (پیوست) آمده است.

از مجموع 4410 ستون محافظت شده که در هم‌ردیفی‌های چندتایی 76 خانواده‌ی پروتئینی در هم‌ردیفی‌های مرجع وجود دارد، 4209 ستون (رزیدوی تشکیل دهنده‌ی موتیف‌های ساختاری) (بیش از 95 درصد) در بزرگترین زیرگراف مشترک حاصل از آخرین هم‌ردیفی حاصل از PSNetal دیده می‌شود. به طور کلی، این عدد گویای این واقعیت است که برنامه‌ی PSNetal در هم‌ردیفی ساختاری چندتایی موفق عمل کرده است. این برنامه، قادر است صرفاً با استفاده از شباهت‌های توپولوژیک و زیستی رزیدوهای آمینو اسیدی در ساختار پروتئین و بدون توجه به توالی آنها، اعضای خانواده‌های مختلف را هم‌ردیف کرده و رأس‌های مشترک میان این ساختارها را استخراج کند.

در 75 درصد از خانواده‌های مطالعه شده (57 خانواده از مجموع کل 76 خانواده) 100 درصد موتیف‌های ساختاری موجود در هر خانواده در انتهای هم‌ردیفی چندتایی بدست می‌آید (شکل 4). نکته‌ی قابل توجه در این زمینه، پراکندگی قابل توجه درصدهای یکسانی توالی، طول پروتئین‌های عضو و تعداد اعضا در این خانواده‌ها است. در حدود 13 درصد (10 خانواده از مجموع کل 76 خانواده) از خانواده‌های مورد بررسی، بیش از 90 درصد موتیف‌های ساختاری موجود در هر خانواده در بزرگترین زیرگراف مشترک حاصل از هم‌ردیفی‌های چندتایی دیده می‌شود (شکل 4). در این خانواده‌ها، در انتهای هم‌ردیفی معمولاً چند رأس از قطعات ساختاری محافظت شده حذف می‌شود که این امر را می‌توان به خطای حاصل از هم‌ردیفی‌های مکرر دوتایی و استخراج گراف جدید از نتیجه‌ی آنها نسبت داد. تنها در 3 خانواده‌ی پروتئینی درصد موتیف‌های ساختاری موجود در زیرگراف مشترک نسبت به موتیف‌های ساختاری بدست آمده در پایگاه داده‌ی MegaMotifBase کمتر از 50 درصد است (نمودارهای 2 و 3). همان‌طور که در بخش نتایج عنوان شد، دو مورد از این سه خانواده دارای اعضایی با طول کوچکتر از 55 آمینو اسید می‌باشند. طول کوچک اعضا و عدم فشردگی و انسجام ساختاری در آنها باعث شده است که رأس‌های مختلف در ساختار فاقد خصوصیات متمایز توپولوژیکی باشد.

با توجه به اصل محافظت‌شدگی بیشتر ساختار نسبت به توالی در تکامل، خانواده‌هایی دیده می‌شوند که در آنها شباهت‌های ساختاری بیش از شباهت در سطح توالی است. خانواده‌ی ribulose bisphosphate carboxylase large chain در این مطالعه، مثالی از این خانواده‌ها است.

بررسی جواب‌های حاصل از برنامه در مقادیر مختلف پارامترهای a وc نشان می‌دهد که هنگامی که مقدار پارامتر a کم (0/1) و پارامتر c بزرگ (0/9) انتخاب شود، درصد درستی یال‌ها افزایش یافته و جواب نهایی بهبود می‌یابد. در مورد خانواده‌ی glycosyl hydrolase family 11 (شکل 6) نیز با تغییر پارامترهای ورودی a و c با الگوی گفته شده در بالا می‌توان جواب‌های حاصل از هم‌ردیفی را به صورت قابل توجهی بهبود بخشید. بنابراین با افزایش وزن ماتریس شباهت توپولوژیک در مواردی‌که درصد یکسانی توالی‌ها کم و میزان شباهت ساختاری زیاد است، می‌توان درصد حضور موتیف‌های ساختاری در بزرگترین زیرگراف مشترک را افزایش داد.

مدت زمان کوتاه انجام محاسبات و سادگی پارامترهای ورودی در برنامه‌ی PSNetal آن را به انتخابی مناسب برای انجام هم‌ردیفی‌های ساختاری چندتایی تبدیل می‌کند. در نسخه‌ی فعلی برنامه که شباهت زیستی شبکه‌های پروتئینی در قالب شباهت اسیدهای آمینه‌ی تشکیل دهنده آنها در ماتریس جایگزینی Blosum62 تعریف شده است، هم‌ردیفی‌های ساختاری چندتایی خانواده‌هایی با شباهت 90-50 درصد با موفقیت انجام می‌شود. در این خانواده‌ها به دلیل تشابه توالی و ساختار میان اعضا، بررسی شباهت میان شبکه‌های ساختاری پروتئین و استخراج بزرگترین زیرگراف مشترک، به تنهایی و با استفاده از ماتریس‌های جایگزینی اسیدهای آمینه (Blosum62) امکان‌پذیر است. با افزایش فاصله‌ی تکاملی میان اعضای یک خانواده و فرایندهایی همانند جهش، حذف و اضافه شدن رزیدوهای آمینو اسیدی رخ می‌دهد که منجر به ایجاد تغییرات موضعی در ساختارها می‌شود. از این رو برای بررسی شباهت‌های ساختاری، استفاده از فاکتور فاصله به‌تنهایی برای یافتن همسایه‌های مجاور و شباهت در سطح اسیدهای آمینه کافی نیست و برای بررسی‌های دقیق‌تر نیاز به بکارگیری خصوصیات ساختاری پیچیده‌تر همانند محتوای ساختار دوم، سطح در دسترس و فاصله‌ی کربن‌های بتا می‌باشد. افزودن این خصوصیات به ماتریس شباهت زیستی به راحتی امکان‌پذیر است و باعث بهبود عملکرد برنامه در هم‌ردیفی‌های ساختاری چندتایی خانواده‌هایی با درصدهای یکسانی کمتر در توالی می‌شود. برنامه‌ی PSNetal در قالب فعلی برای انجام هم‌ردیفی‌های ساختاری سرتاسری بهینه شده است. مرحله‌ی بعدی تحقیقات، تهیه‌ی نسخه‌ای از برنامه‌ی PSNetal است که قادر به انجام هم‌ردیفی‌های موضعی چندتایی بین اعضای خانواده‌های پروتئینی باشد و جایگاه‌های فعال آنزیم را در هر خانواده شناسایی و استخراج کند.

جدول 1 خانواده‌های پروتئینی استخراج شده از پایگاه HOMSTRAD و بررسی نتایج همردیفی چندتایی از لحاظ تشخیص موتیف های ساختاری. در این جدول، ستون هفتم تعداد قطعات یا بلوک‌های ساختاری محافظت شده و تعداد رزیدوهای تشکیل دهنده‌ی آن در هر خانواده در پایگاه (MegaMotifBase(MMB، ستون هشتم تعداد قطعات یا بلوک‌های ساختاری محافظت شده در هر خانواده در نتایج حاصل از اجرای برنامه‌ی(PSNetAl (PSN، را نشان می‌دهد.

ردیف	نام خانواده	درصد یکسانی توالی‌ها	اندازهٔ متوسط ساختارها	تعداد اعضا	درصد درستی یال‌ها	تعداد موتیف‌های ساختاری در MMB	تعداد موتیف‌های ساختاری در PSN	درصد موتیف‌های موجود در هر خانواده در PSN
1	metallothionine_alpha domain	90	31	3	0/88	16	16	%100
2	Paralytic/GBP/PSP peptide	83	23	3	0/9	6	5	%83/33
3	Transthyretin	82	116	3	0/93	23	23	%100
4	cytokine -- granulocyte colony stimulating factor	82	153	3	0/96	59	59	%100
5	metallothionine_beta domain	80	30	3	0/803	20	20	%100

6	B domain	78	50	3	0/88	20	20	%100
7	Pyridoxal-dependent decarboxylase, C-terminal sheet domain	78	152	3	0/95	39	39	%100
8	histocompatibility antigen-binding domain	76	178	5	0/95	40	40	%100
9	Photosystem I reaction centre subunit IV / PsaE	74	69	3	0/91	6	6	%100
10	pancreatic lipase	74	446	5	0/92	141	141	%100
11	copper-containing nitrite reductase	73	334	3	0/97	104	104	%100
12	Methyl-coenzyme M reductase alpha subunit, C-terminal domain	71	279	3	0/97	85	85	%100
13	heat-labile enterotoxin, A subunit	70	187	3	0/93	51	51	%100
14	Methyl-coenzyme M reductase beta subunit, C-terminal domain	69	252	3	0/95	99	99	%100
15	xylose isomerase	69	388	6	0/97	119	119	%100
16	Ubiquitin homologues	68	74	3	0/96	20	20	%100
17	fructose-1,6-bisphosphatase	68	321	4	0/98	94	94	%100
18	Nitric oxide synthase, oxygenase domain	67	386	3	0/93	114	114	%100
19	DHH	67	185	3	0/94	62	62	%100
20	Methyl-coenzyme M reductase alpha subunit	67	550	3	0/96	164	164	%100
21	cold-shock DNA-binding domain	67	67	3	0/99	18	18	%100
22	Antifreeze protein	66	67	3	0/89	26	26	%100
23	Nickel-dependent hydrogenases, large subunit	66	535	3	0/96	142	142	%100
24	cyclin-dependent kinases regulatory subunit	65	81	3	0/95	9	9	%100
25	alcohol dehydrogenase	64	373	5	0/93	117	106	%90/59
26	DHHA2 domain	63	119	3	0/88	36	34	%94/44
27	Methyl-coenzyme M reductase alpha subunit, N-terminal domain	63	271	3	0/97	55	55	%100
28	Cyclodextrin glycosyltransferase	63	684	6	0/98	168	168	%100
29	rubredoxin	62	51	5	0/86	19	9	%47/36
30	transferrin	62	526	7	0/91	103	103	%100
31	Methyl-coenzyme M reductase beta subunit	62	436	3	0/94	174	174	%100
32	TATA-box binding protein, C-terminal domain	61	183	4	0/92	62	62	%100
33	Microbial ribonucleases	61	104	3	0/92	23	23	%100
34	fructose-1,6-bisphosphate aldolase	61	355	3	0/93	109	109	%100
35	Bulb-type mannose-specific lectin	60	108	3	0/92	22	22	%100
36	Methyl-coenzyme M reductase gamma subunit	60	247	3	0/94	74	74	%100
37	NAD(P)H dehydrogenase (quinone)	60	258	3	0/95	101	101	%100
38	ferritin	60	166	4	0/98	60	60	%100
39	Chitin binding domain	59	42	5	0/97	13	13	%100
40	Colicin immunity protein / pyocin immunity protein	58	85	3	0/93	18	18	%100
41	snake venom metalloproteinase	58	199	3	0/93	65	65	%100
42	nerve growth factor	57	109	4	0/92	15	15	%100
43	Thaumatin	57	207	3	0/94	63	63	%100
44	annexin	57	317	6	0/94	93	47	%50/53
45	NADH ubiquinone oxidoreductase, 20 Kd subunit	57	265	5	0/96	65	65	%100
46	calcium-binding protein -- parvalbumin-like	56	107	7	0/89	27	27	%100
47	Adenylosuccinate synthetase	56	430	3	0/91	152	149	%98/026
48	nucleotide diphosphate kinase	56	149	4	0/95	47	47	%100

49	IPT/TIG domain	55	84	6	0/93	23	23	%100
50	glyceraldehyde 3-phosphate dehydrogenase	55	388	8	0/95	74	72	%97/29
51	matrix metalloproteinase	55	164	6	0/96	57	57	%100
52	Nucleotidyltransferase, domain 2	54	145	3	0/88	24	23	%95/83
53	Low molecular weight phosphatase	54	157	3	0/92	53	46	%86/79
54	Elongation factor Tu (EF-Tu), C-terminal domain	54	97	3	0/98	28	28	%100
55	Domain containing Gla	53	40	4	0/73	7	2	%28/57
56	hormone receptor (DNA-binding domain)	53	74	5	0/93	13	13	%100
57	Zinc-binding domain present in Lin-11, Isl-1, Mec-3.	53	69	5	0/96	22	11	%50
58	thionin	53	46	3	0/98	10	10	%100
59	Starch binding domain	52	105	8	0/78	19	12	%63/15
60	Adenovirus fiber protein head domain (knob domain)	52	188	3	0/86	66	61	%92/42
61	Fork head domain	52	92	3	0/87	10	10	%100
62	glutaminase-asparagainase	52	326	4	0/90	105	105	%100
63	heterotrimeric G proteins - alpha subunit	52	317	3	0/90	98	94	%95/918
64	YgbB family	52	153	3	0/90	50	48	%96
65	Methyl-coenzyme M reductase beta subunit, N-terminal domain	52	184	3	0/91	43	43	%100
66	Macrophage migration inhibitory factor (MIF)	52	115	3	0/93	29	29	%100
67	immunoglobulin domain -- V set - immunoglobulin heavy chain	52	123	21	0/94	25	24	%96
68	serine proteinase -- bacterial	52	188	5	0/96	59	59	%100
69	phosphoglycerate kinase	51	405	4	0/93	126	126	%100
70	glycosyl hydrolase family 22 (lysozyme)	50	126	12	0/95	24	23	%95/83
71	FERM domain (Band 4.1 family), region 2	50	109	3	0/86	16	16	%100
72	glycosyl hydrolase family 11	50	185	5	0/62	57	0	0
73	Ribulose bisphosphate carboxylase large chain	50	453	7	0/83	111	86	%77/48
74	cytochrome-c5	50	81	6	0/86	7	7	%100
75	immunoglobulin domain -- V set	50	112	26	0/91	29	29	%100
76	legume lectin	50	234	11	0/97	67	67	%100

6- منابع

[1] Krissinel, E. and K. Henrick, *Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions.* Acta Crystallographica Section D: Biological Crystallography, 2004. **60**(12): pp. 2256-2268.

[2] Konagurthu, A.S., et al, *MUSTANG: A Multiple Structural Alignment Algorithm.* PROTEINS: Structure, Function, and Bioinformatics, 2006. **64**: pp. 559–574.

[3] Kolodny, R. and N. Linial, *Approximate Protein Structural Alignment in Polynomial Time.*

[4] Krissinel, E. and K. Henrick, *Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions.* Acta Crystallographica Section D, 2004: pp. 2256–2268.

[5] Shealy, P. and H. Valafar, *Multiple structure alignment with msTALI.* BMC Bioinformatics, 2012. **13.** (105)

[6] Mayr, G., F.S. Domingues, and P. Lackner, *Comparative analysis of protein structure alignments.* BMC Struct Biol, 2007. **7**: pp. 50.

[7] Shindyalov, I.N. and P.E. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.* Protein engineering, 1998. **11**(9): pp. 739-747.

[8] Holm, L. and C. Sander, *Protein structure comparison by alignment of distance matrices.* Journal of molecular biology, 1993. **233**(1): pp. 123-138.

[9] Ye, Y. and A. Godzik, *FATCAT: a web server for flexible structure comparison and structure similarity searching.* Nucleic acids research, 2004. **32**(suppl 2): pp. W582-W585.

[10] Ye, Y. and A. Godzik, Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics, 2003. 19 (suppl 2): pp. ii246-ii255.

[1] Jung, J. and B. Lee, *Protein structure alignment using environmental profiles.* Protein Engineering, 2000. **13**(8): pp. 535-543.

[2] Kawabata, T., *MATRAS: a program for protein 3D structure comparison.* Nucleic acids research, 2003. **31**(13): pp. 3367-3369.

[3] Ma, J. and S. Wang, *Algorithms, applications, and challenges of protein structure alignment.* Adv. Protein Chem. Struct. Biol, 2014. **94**: pp. 121-175.

[4] Shealy, P. and H. Valafar, *Multiple structure alignment with msTALI.* BMC bioinformatics, 2012. **13**(1): pp. 105.

[5] Menke, M., B. Berger, and L. Cowen, *Matt: local flexibility aids protein multiple structure alignment.* PLoS Comput Biol, 2008. **4**(1): pp. e10.

[6] Konagurthu, A.S., et al., *MUSTANG: a multiple structural alignment algorithm.* Proteins: Structure, Function, and Bioinformatics, 2006. **64**(3): pp. 559-574.

[7] Guda, C., et al., *CE-MC: a multiple protein structure alignment server.* Nucleic acids research, 2004. **32**(suppl 2): pp. W100-W103.

[8] Neyshabur, B., et al., *NETAL: a new graph-based method for global alignment of protein–protein interaction networks.* Bioinformatics, 2013. **29**(13): pp. 1654-1662.

[9] da Silveira, C.H., et al., *Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins.* Proteins: Structure, Function, and Bioinformatics, 2009. **74**(3): pp. 727-743.

[10] Singh, R., J. Xu, and B. Berger, *Global alignment of multiple protein interaction networks with application to functional orthology detection.* Proceedings of the National Academy of Sciences, 2008. **105**(35): pp. 12763-12768.

[11] Kuchaiev, O., et al., *Topological network alignment uncovers biological function and phylogeny.* Journal of the Royal Society Interface, 2010. **7**(50): pp. 1341-1354.

[12] Slater, A.W., et al., *Towards the development of standardized methods for comparison, ranking and evaluation of structure alignments.* Bioinformatics, 2013. **29**(1): pp. 47-53.

[13] Zhu, J. and Z. Weng, *FAST: a novel protein structure alignment algorithm.* PROTEINS: Structure, Function, and Bioinformatics, 2005. **58**(3): pp. 618-627.

[14] Kim, C. and B. Lee, *Accuracy of structure-based sequence alignment of automatic methods.* BMC bioinformatics, 2007. **8**(1): p. 355.

[15] Pugalenthi, G., et al., *MegaMotifBase: a database of structural motifs in protein families and superfamilies.* Nucleic acids research, 2008. **36**(suppl 1): pp. D218-D221.

[16] Stebbings, L.A. and K. Mizuguchi, *HOMSTRAD: recent developments of the homologous protein structure alignment database.* Nucleic acids research, 2004. **32**(suppl 1): pp. D203-D207.

[17] Sierk, M.L. and W.R. Pearson, *Sensitivity and selectivity in protein structure comparison.* Protein Science, 2004. **13**(3): pp. 773-785.

[18] Koehl, P., *Protein structure similarities.* Current opinion in structural biology, 2001. **11**(3): pp. 348-353.

[19] Berbalk, C., C.S. Schwaiger, and P. Lackner, *Accuracy analysis of multiple structure alignments.* Protein Science, 2009. **18**(10): pp. 2027-2035.