ارائه روشی برای تطبیق چندگانه دنباله*های پروتئین با استفاده از الگوریتم‌های ژنتیکی

زهرا نریمانی ، حسن ابوالحسنی

دانشکده مهندسی کامپیوتر دانشگاه صنعتی شریف، تهران، ایران
narimani@ce.sharif.edu, abolhassani@sharif.edu

چکیده

تطبیق چندگانه دنباله یکی از مسائل مهم در زیست شناسی است که الگوریتم چندجمله‌ای برای حل آن وجود ندارد. با توجه به اهمیت این مسئله تا کنون روش‌های زیادی برای یافتن راه حل‌های نسبتاً دقیق برای این مسئله ارائه شده است. از این قبیل راهکارها می‌توان به استفاده از الگوریتم‌های تقریبی، استفاده از الگوریتم‌های تکاملی، و استفاده از مدل‌های احتمالاتی اشاره کرد. الگوریتم‌های تکاملی و به طور خاص الگوریتم‌های ژنتیکی با توجه به خصوصیات خاص آنها، می‌توانند جزء راهکارهای مناسب حل این مسئله باشند. در این مقاله الگوریتم‌های ژنتیکی ارائه شده برای حل این مسئله را بررسی خواهیم کرد و با اشاره به معایب هر یک، روشی را ارائه می‌دهیم که تا حد امکان مشکلات این روش‌ها را بهبود بخشد. یکی از این مشکلات، وابستگی به داشتن اطلاعاتی از قبیل میزان شباهت دنباله‌های مورد تطبیق، برای انتخاب ماتریس جانشینی مناسب، است که در این مقاله با روشی تقریبی این نیاز برطرف شده است. آزمایشات برای بررسی میزان دقت الگوریتم بر روی مجموعه داده‌ی BAliBASE2.0، که یک مجموعه‌ی استاندارد از دنباله‌های منطبق شده می‌باشد، انجام شده است. با توجه به اینکه تطبیق چندگانه‌ی رشته‌هایی که شباهت آنها با یکدیگر کمتر است، مسئله‌ی پیچیده‌تری می‌باشد در این مقاله از زیرمجموعه‌هایی از مجموعه داده‌ی مورد نظر استفاده شده است که شباهت کمی (کمتر از ۲۵٪) با یکدیگر دارند. نتایج نشان می‌دهد که الگوریتم ارائه شده با استفاده از روش خاص در مقدار دهی اولیه جمعیت و استفاده از عملگرهای ساده و دقیق، دقت بهتری را پس از تعداد کمی نسل ژنتیکی نسبت به روش‌های مشابه نتیجه داده است. در واقع رسیدن به یک پاسخ مناسب پس از ایجاد تعداد کمی نسل ژنتیکی (حدود ۲۰۰۰) یکی دیگر از مشکلات الگوریتم‌های ژنتیکی موجود که همگرایی کند آنهاست را جبران می‌نماید.

کلمات کلیدی

تطبیق چندگانه دنباله، الگوریتم‌های ژنتیکی، ماتریس جانشینی

* Multiple sequence alignment

۱ مقدمه

مطالعه بر روی ساختارهای زیستی مانند پروتئین، RNA و DNA از مسائل مهم در بیولوژی است. به دلیل پیچیده و طویل بودن ساختارهای زیستی و همچنین بزرگ بودن دیتاست‌های مورد مطالعه، برای بررسی خواص این نوع داده به الگوریتم‌های کارایی نیاز داریم. یکی از شاخه‌های این مطالعات، الگوریتم‌هایی هستند که برای بررسی میزان شباهت و تطبیق این دنباله‌ها به کار می‌روند. مسئله تطبیق دنباله در مواردی مانند تشکیل درخت تکامل نژادی جانداران، و در پیشگویی عمل و ساختار پروتئین‌های ناشناخته از روی ساختار و عملیات پروتئین‌های شناخته شده، و یافتن نواحی با عملکرد خاص در دنباله‌های پروتئین کاربرد دارد. از چنین اطلاعاتی می‌توان در مواردی مانند طراحی داروها یا بهبود داروهای موجود، و کمک به پیش بینی اطلاعات ساختاری پروتئین‌ها استفاده کرد.

مسئله‌ی تطبیق دنباله معمولاً به صورت یک مسئله‌ی بهینه سازی تعریف می‌شود که در آن هدف یافتن تطبیقی با بیشترین امتیاز می‌باشد. نحوه‌ی امتیاز دهی به تطبیق‌ها بر اساس تعداد و نوع آمینو اسیدهای منطبق شده بر یکدیگر محاسبه می‌شود. این امتیازها در ماتریس‌هایی به نام ماتریس‌های جانشینی تعریف می‌شود. استفاده از این شیوه‌ی امتیازدهی علیرغم اینکه، با مشخص کردن میزان کیفیت یک تطبیق، به یافتن پاسخ بهینه کمک می‌کند دارای مشکلاتی نیز می‌باشد؛ برای مثال استفاده از این ماتریس‌ها نیاز به داشتن اطلاعاتی از قبیل فاصله‌ی تکاملی[۲] دنباله‌های مورد نظر تا نزدیکترین جد مشترک دنباله‌ها می‌باشد و مشکل دیگر این است که به دلیل ناشناخته بودن واقعیات مربوط به تکامل، این ماتریس‌ها خود از روش‌های احتمالاتی تولید شده‌اند و نادقیق هستند. عدم وجود الگوریتمی چند جمله‌ای برای تطبیق چندگانه‌ی رشته و عدم قطعیت در مورد اطلاعات مورد نیاز برای انجام یک تطبیق، یافتن پاسخ مناسبی برای این مسئله را مشکل می‌سازد. در این مقاله با اشاره به راهکارهای موجود برای تطبیق چندگانه، روشی با استفاده از الگوریتم‌های ژنتیکی برای آن ارائه خواهد شد.

۲ تطبیق دنباله‌های پروتئین

مسئله‌ی تطبیق می‌تواند بین دو دنباله مطرح شود. در این حالت که تطبیق دوگانه[۳] نام دارد، مسئله شبیه به همان مسئله‌ی ساده‌ی یافتن کمترین میزان تغییرات ممکن برای تبدیل یک دنباله به یک دنباله‌ی دیگر است که یکی از الگوریتم‌های اولیه برای حل آن، Needlman-Waunch [۱۹] است که از برنامه سازی پویا استفاده می‌کند. در تطبیق چندگانه، تعداد رشته‌ها بیشتر از دو است. حتی مسئله‌ی تطبیق دوگانه نیز در مورد رشته‌های زیستی به همین سادگی نیست و همیشه نمی‌توان کاراکترهای مشابهی را برای منطبق کردن یافت. بلکه کاراکترهای متفاوت هم به شرطی که از نظر زیستی احتمال تبدیل آنها به داشته یکدیگر وجود باشد، قابل منطبق شدن بر یکدیگر هستند. در واقع ماتریس‌هایی به نام ماتریس‌های جانشینی وجود دارد که برای هر دو کاراکتر در رشته‌های زیستی (که در DNA نوکلئوتید و در پروتئین اسید آمینه نام دارند) میزان احتمال تبدیل آنها به یکدیگر را مشخص می‌کند. این احتمالات به صورت امتیازهایی در این ماتریس‌ها تعریف شده‌اند.

تعریف رسمی: برای تعریف رسمی مسئله از تعریفی که در [۱۸] ارائه شده است استفاده می‌کنیم. فرض کنیم که $S=\{S_1, S_2, \ldots, S_n\}$ مجموعه‌ای از n رشته باشد که روی الفبای A تعریف شده‌اند، و هر رشته‌ی S_I در آن شامل l_i کاراکتر دارای ترتیب مشخص باشد:

$$S_i = s_{i1}s_{i2}....s_{il_i}, \qquad \forall i = 1,2,...,n$$

الفبای جدید $\hat{A} = A \cup \{-\}$ را با اضافه کردن کاراکتر فاصله[۴] به الفبای موجود تعریف می‌کنیم. مجموعه‌ی $\hat{S} = \{\hat{S}_1, \hat{S}_2, ..., \hat{S}_n\}$ که مجموعه‌ای از رشته‌های تعریف شده روی الفبای $\hat{A}$ است را بعنوان تطبیق رشته‌های مجموعه‌ی S تعریف می‌کنیم اگر شرایط زیر برقرار باشد:

۱. تمام رشته‌های موجود در $\hat{S}$ طول مساوی و برابر $\hat{l}$ داشته باشند، و داشته باشیم: $\max_{i=1...n}(l_i) \le \hat{l} \le \sum_{i=1}^{n} l_i$

۲. با در نظر نگرفتن کاراکترهای فاصله، هر $\hat{S}_i$ برابر با رشته‌ی S_i باشد (برای هر $i = 1,2,..., n$).

۳. $\hat{S}$ دارای هیچ ستونی که تنها شامل کاراکتر فاصله است، نباشد.

بنابراین یک تطبیق چندگانه را می‌توان به صورت ماتریسی n سطری در نظر گرفت، به طوریکه سطر i ام در آن شامل $\hat{S}_i$ است. در شکل ۱ نمونه‌ای از تطبیق دوگانه و چندگانه را مشاهده می‌کنید.

تطبیق دوگانه

1	C A - T G A G - A T C
2	- A C T C A G T A - C

تطبیق چندگانه

1	C A - T G A G - A T C
2	- A C T C A G T A - C
3	C A - A G - G - A T C
4	G - - T C A G T A - C

شکل ۱- نمونه‌ای از تطبیق دوگانه و تطبیق چندگانه

[۲] evolutionary distance

[۳] Pairwise alignment

[۴] gap

برای حل مسئله‌ی تطبیق، نیاز داریم به نحوی میزان خوب بودن یک تطبیق را به دست آوریم. یک رویکرد ساده، جریمه کردن میزان تفاوت دو رشته بر حسب اضافه[۵]، حذف[۶] یا جایگزینی[۷] کاراکترها می‌باشد. یک تطبیق مناسب از نظر تئوری، تطبیقی است که بیشترین امتیاز را از نظر جفت کاراکترهای منطبق شده با یکدیگر به دست بدهد. در [۱]، [۲] و [۹] می‌توان نمونه‌هایی از امتیازدهی را مشاهده کرد.

۳ ماتریس‌های جانشینی

برای اینکه مشخص کنیم هر دو کاراکتر زیستی با چه امتیازی بر یکدیگر منطبق می‌شوند از ماتریس‌های جانشینی استفاده می‌کنیم. در واقع هر عدد در این ماتریس‌ها می‌تواند بر اساس روش‌های خاصی، مانند بررسی و مقایسه‌ی خواص فیزیکی و شیمیایی کاراکترها، استفاده از مدل‌های تکاملی و روش‌های آماری برای بررسی میزان احتمال تبدیل کاراکترها به یکدیگر در فرایند تکامل، و تنظیم دستی امتیازها توسط افراد خبره، تولید شود.

مجموعه ماتریس‌های PAM و BLOSUM نمونه‌ای از ماتریس‌های جانشینی هستند که با استفاده از روش‌های آماری به دست آمده‌اند. سری BLOSUM به دلیل دقت بیشتر در مدل سازی اولیه نتایج بهتری در بر داشته و پرکاربردتر نیز می‌باشد. برای استفاده از این ماتریس‌ها باید اطلاعاتی در مورد رشته‌ها داشته باشیم. برای مثال برای استفاده از ماتریس BLOSUM باید میزان شباهت بین رشته‌های مورد نظر را بدانیم. در شکل ۲ ماتریس BLOSUM80 را مشاهده می‌کنید که برای جانشینی اسیدآمینه‌ها در پروتئین‌هایی با ۸۰٪ شباهت استفاده می‌شود.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	7	-3	-3	-3	-1	-2	-2	0	-3	-3	-3	-1	-2	-4	-1	2	0	-5	-4	-1
R	-3	9	-1	-3	-6	1	-1	-4	0	-5	-4	3	-3	-5	-3	-2	-2	-5	-4	-4
N	-3	-1	9	2	-5	0	-1	-1	1	-6	-6	0	-4	-6	-4	1	0	-7	-4	-5
D	-3	-3	2	10	-7	-1	2	-3	-2	-7	-7	-2	-6	-6	-3	-1	-2	-8	-6	-6
C	-1	-6	-5	-7	13	-5	-7	-6	-7	-2	-3	-6	-3	-4	-6	-2	-2	-5	-5	-2
Q	-2	1	0	-1	-5	9	3	-4	1	-5	-4	2	-1	-5	-3	-1	-1	-4	-3	-4
E	-2	-1	-1	2	-7	3	8	-4	0	-6	-6	1	-4	-6	-2	-1	-2	-6	-5	-4
G	0	-4	-1	-3	-6	-4	-4	9	-4	-7	-7	-3	-5	-6	-5	-1	-3	-6	-6	-6
H	-3	0	1	-2	-7	1	0	-4	12	-6	-5	-1	-4	-2	-4	-2	-3	-4	3	-5
I	-3	-5	-6	-7	-2	-5	-6	-7	-6	7	2	-5	2	-1	-5	-4	-2	-5	-3	4
L	-3	-4	-6	-7	-3	-4	-6	-7	-5	2	6	-4	3	0	-5	-4	-3	-4	-2	1
K	-1	3	0	-2	-6	2	1	-3	-1	-5	-4	8	-3	-5	-2	-1	-1	-6	-4	-4
M	-2	-3	-4	-6	-3	-1	-4	-5	-4	2	3	-3	9	0	-4	-3	-1	-3	-3	1
F	-4	-5	-6	-6	-4	-5	-6	-6	-2	-1	0	-5	0	10	-6	-4	-4	0	4	-2
P	-1	-3	-4	-3	-6	-3	-2	-5	-4	-5	-5	-2	-4	-6	12	-2	-3	-7	-6	-4
S	2	-2	1	-1	-2	-1	-1	-1	-2	-4	-4	-1	-3	-4	-2	7	2	-6	-3	-3
T	0	-2	0	-2	-2	-1	-2	-3	-3	-2	-3	-1	-1	-4	-3	2	8	-5	-3	0
W	-5	-5	-7	-8	-5	-4	-6	-6	-4	-5	-4	-6	-3	0	-7	-6	-5	16	3	-5
Y	-4	-4	-4	-6	-5	-3	-5	-6	3	-3	-2	-4	-3	4	-6	-3	-3	3	11	-3
V	-1	-4	-5	-6	-2	-4	-4	-6	-5	4	1	-4	1	-2	-4	-3	0	-5	-3	7

شکل ۲ - ماتریس Blosum80 [۱۷] برای جانشینی اسید آمینه‌ها در پروتئین‌هایی با ۸۰ درصد شباهت تعریف شده است.

این که چگونه میزان شباهت رشته‌ها را قبل از انجام تطبیق چندگانه به دست آوریم خود یک مسئله است که با رویکردهای متفاوتی مورد بررسی قرار می‌گیرد.

۴ کارهای انجام شده

روش‌های کلی موجود برای تطبیق چندگانه را می‌توان به دسته‌های دقیق، روش‌های پیشرفتی، روش‌های تکراری، و روش‌های آماری دسته بندی کرد. در روش‌های دقیق سعی می‌شود یک تطبیق بهینه (زیربهینه) برای رشته‌ها پیدا شود. دقیق بودن این الگوریتم‌ها باعث می‌شود تعداد رشته ای که این روش‌ها قادر به حل مسئله‌ی تطبیق در مورد آنها شوند و همچنین تابع امتیازدهی‌ای که قادر به بهینه کردن آن هستند نیز محدود شوند. در واقع استفاده از الگوریتم دقیق معمولاً فقط در مورد دو رشته انجام می‌شود. در روش‌های پیشرفتی با گسترش الگوریتم تطبیق دو رشته، تطبیق چندگانه نیز انجام می‌شود. به این منظور معمولاً از یک هسته‌ی اولیه، که یک تطبیق دوگانه در میان کل رشته‌هاست و ممکن است با معیار خاصی انتخاب شود، استفاده شده و سپس سعی می‌شود سایر رشته‌ها یکی یکی به این تطبیق اضافه شوند. الگوریتم‌های T-Coffee [۸]، MultAlign [۱۴]، و ClastalW [۴] نمونه‌هایی از الگوریتم‌هایی هستند که به صورت پیشرفتی عمل می‌کنند. مزیت این روش‌ها این است که به سرعت به جواب می‌رسند. بزرگترین مشکل این الگوریتم‌ها این است که اگر هسته‌ی اولیه به درستی انتخاب نشود، تطبیق نهایی کاملاً منحرف شده و به پاسخ مناسبی نمی‌رسد. در روش‌های تکراری سعی می‌شود طی عملیاتی تکرار شونده تطبیق‌ها تشکیل و بهبود داده شوند. روش-هایی که از الگوریتم‌های تکاملی استفاده می‌کنندو همچنین روش‌هایی که از simulated annealing استفاده می‌کنند از جمله این روش‌ها هستند. روش‌های SAGA [۹]، AMPS، Praline [۱۵]، و IterAlign [۱۶] الگوریتم‌های تکراری برای حل مسئله‌ی تطبیق چندگانه هستند. این روش‌ها چندان سریع نیستند و تضمینی برای پیدا کردن جواب بهینه ندارند ولی ماهیت تکراری آنها باعث می‌شود انعطاف بیشتری نسبت به روش‌های پیشرفتی در یافتن جواب از خود نشان دهند. دسته‌ای از الگوریتم‌های تطبیق بر اساس روش‌های آماری مانند مدل پنهان مارکوف[۸] استفاده می‌کنند. در این روش‌ها سعی می‌شود برای رشته‌های موجود یک مدل احتمالی (پروفایل) تشکیل شود و رشته‌ها با توجه به این مدل با یکدیگر تطبیق داده شوند. از مشکلات این روش‌ها نیاز به تعداد زیادی رشته در ورودی برای تشکیل یک مدل احتمالی درست و همچنین نیاز به در نظر گرفتن پیش فرض‌هایی که غلط بودن آنها تطبیق را منحرف می‌کند، می‌باشد.

بررسی‌های انجام شده نشان می‌دهد روش‌های تکراری مانند الگوریتم-های ژنتیکی از انعطاف بیشتری در مورد تابع امتیاز دهی مورد استفاده

[۵] insertion

[۶] deletion

[۷] substitution

[۸] Hidden Markov Model

برخوردارند و مقالات اخیری که با استفاده از این روش‌ها تطبیق چندگانه را انجام داده‌اند نسبت به سایر الگوریتم‌های ارائه شده دقت بیشتری ارائه کرده‌اند. مشکلی که این روش‌ها دارند کندی آنها و عدم وجود تضمین برای یافتن پاسخ بهینه است. در این مقاله روش‌های مبتنی بر الگوریتم‌های ژنتیکی مدنظر قرار داده شده است و روشی با این روش ارائه شده است که در آن سعی شده است رفع مشکلات رایج موجود در این روش‌ها مدنظر قرار داده شود.

SAGA [۹] یکی از اولین الگوریتم‌های ژنتیکی برای حل این مسئله است که از ۲۲ عملگر برای تولید نسل جدید از نسل قبلی استفاده می‌کند. این موضوع باعث پیچیده شدن این روش از لحاظ تنظیم پارامترها و زمان اجرا می‌شود. در SAGA احتمال وقوع همه‌ی عملگرها در ابتدا با یکدیگر مساوی است ولی با روند الگوریتم و با بررسی اثر هر عملگر، این احتمالات بطور خودکار تنظیم می‌شوند. روش دیگری در [۲] ارائه شده است که از الگوریتم‌های ژنتیکی برای بهبود تطبیق به دست آمده از ClustalV [۱۳] استفاده می‌کند. این روش دارای پنج عملیات ژنتیکی تعریف شده است. در هر دوی این روش‌ها به این دلیل که جمعیت اولیه به صورت تصادفی به وجود آمده است، ارزیابی تابع امتیازدهی باید به تعداد زیادی روی جمعیت اعمال شود و یک پاسخ مناسب پس از نسل‌های زیادی ایجاد می‌شود. همچنین در این روش‌ها باید ماتریس جانشینی مرتبط در ابتدا توسط کاربر مشخص شود که اینها از نقاط ضعف این دو روش می‌باشند. در [۳] نیز روشی ارائه شده است که عملیاتی تقریباً شبیه به SAGA داشته ولی از روشی موازی برای اجرای الگوریتم ژنتیکی استفاده می‌کند. در این مقاله نشان داده شده است که پاسخی که با این روش موازی به دست آمده است نسبت به پاسخ ClustalW و SAGA بهبود یافته است.

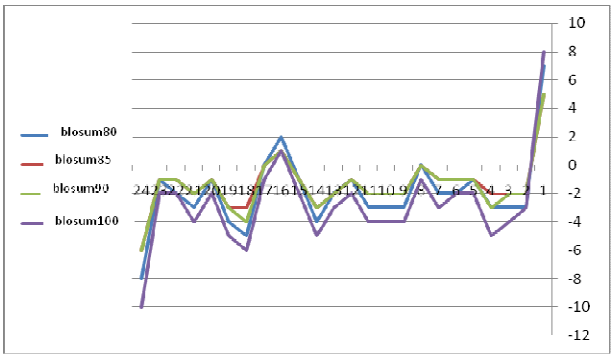
یکی از بهترین روش‌های ارائه شده در این زمینه روشی است که در [۱] بیان شده است. در این روش عملگرها بسیار ساده هستند و همین امر باعث شده است که از نظر زمان اجرا انعطاف بیشتری در الگوریتم پدید آید و رشته‌های طولانی‌تر بتوانند در زمان کمتری با یکدیگر تطبیق داده شوند. مزیت دیگری که این روش دارد مقدار اولیه دهی خاصی است که برای جمعیت ژنتیکی در آن انجام شده است که منجر به همگرایی سریعتر الگوریتم و در نتیجه نیاز به ایجاد تعداد کمتری نسل برای رسیدن به یک پاسخ مناسب شده است. در واقع این الگوریتم به جای استفاده از نمایشی تصادفی برای هر رشته، از یکی از تطبیق‌های دوگانه‌ی آن با رشته‌های دیگر استفاده می‌شود. در ایجاد این تطبیق‌های دوگانه از روش Needlman-Waunch استفاده می‌شود. مشکلی که این روش دارد وابستگی شدید به جمعیت اولیه و عدم استفاده از جریمه‌ی فاصله در تابع امتیاز دهی می‌باشد.

۵ روش ارائه شده

روش جدید ارائه شده در این مقاله، شامل دو مرحله‌ی اصلی می‌باشد. در مرحله‌ی اول ماتریس جانشینی مناسب برای تطبیق رشته‌ها به صورت خودکار انتخاب می‌گردد و در مرحله‌ی دوم الگوریتم ژنتیکی با رویکرد جدیدی بر روی رشته‌ها اجرا می‌گردد. در بخش‌های بعدی هر یک از این مراحل به طور مفصل توضیح داده می‌شود.

۱.۵ تعیین خودکار ماتریس جانشینی

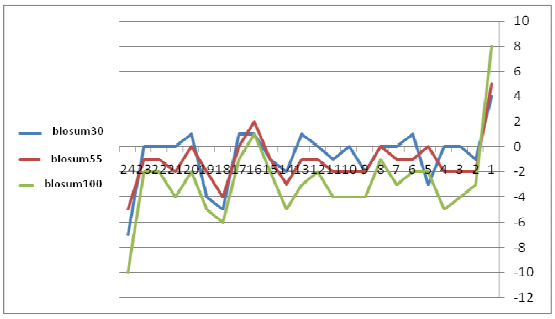
در این مقاله تمرکز ما بر روی ماتریس‌های BLOSUM می‌باشد. برای اینکه بتوانیم روشی برای انتخاب ماتریس مناسب برای تطبیق رشته‌ها پیدا کنیم، ابتدا بر روی داده‌های این ماتریس‌ها تحلیل روند[۹] انجام داده‌ایم. به این صورت که به ازای هر جدول جانشینی اعداد مربوط به تبدیل هر اسید آمینه به تمام اسید آمینه‌های دیگر را به صورت یک نمودار رسم کردیم. در نتیجه برای هر ماتریس ۲۰ نمودار به دست آمد که نشان می‌داد اسید آمینه‌ی مرتبط با هر یک از اسید آمینه‌های دیگر با چه امتیازهایی منطبق می‌شود. برای مثال در جدول BLOSUM80 برای اسید آمینه‌ی A (Alanine) چارت آبی رنگ در نمودار شماره ۱ را داریم.



نمودار ۱- مقایسه‌ی امتیاز تطبیق اسید آمینه‌ی Alanine با اسید آمینه‌های دیگر در بازه‌ی ۸۰-۱۰۰ درصد شباهت

در مقایسه‌ی این نمودارها به این نتیجه رسیدیم که روند صعودی و نزولی این نمودارها در بازه‌های ۳۰-۴۵، ۴۵-۶۰، ۶۰-۸۰ و ۸۰-۱۰۰ تقریباً یکسان است. در نمودار ۱ نمونه‌ای از مقایسه‌ی روند را در بازه‌ی ۸۰-۱۰۰ و در نمودار ۲ نمونه‌ای از این مقایسه را در بازه‌های متفاوت مشاهده می‌کنید.

[۹] Trend Analysis



نمودار ۲- مقایسه‌ی امتیاز تطبیق اسید آمینه‌ی Alanine با اسید آمینه‌های دیگر در بازه‌های متفاوت

همینطور که مشاهده می‌کنید در نمودار ۱ روندها تقریباً یکسان هستند ولی در نمودار ۲ تفاوت‌هایی در این روندها، مثلاً در نقاط ۳، ۵ ، ۷ و ۱۰، دیده می‌شود. با در نظر گرفتن سایر اسید آمینه‌ها این تفاوت‌ها در بازه‌های متفاوت بیشتر نمایان می‌گردد.

با توجه به این مشاهده، راهکاری که در اینجا برای انتخاب ماتریس جانشینی ارائه می‌شود این است که ابتدا رشته‌ها را توسط الگوریتم پیشرفتی و سریع ClustalW توسط ماتریس‌های BLOSUM30، BLOSUM45، BLOSUM60، و BLOSUM80 منطبق می‌کنیم. پس از تطبیق رشته‌ها، میزان شباهت (درصد کاراکترهای یکسانی که زیر هم قرار گرفته‌اند) آنها را در هر یک از تطبیق‌های مذکور به دست می‌آوریم. اگر این درصدها برابر بودند، ماتریس جانشینی متناسب با توجه به درصد شباهت انتخاب می‌کنیم. اگر این درصدها متفاوت بودند، ماتریس جانشینی را انتخاب می‌کنیم که بیشترین درصد شباهت را به دست داده است. نتایج نشان می‌دهد این روش، ماتریس مناسبی را حداقل به طور تقریبی انتخاب می‌کند و تطبیق مناسبی با انتخاب این ماتریس به دست می‌آید.

۲.۵ الگوریتم ژنتیکی تطبیق چندگانه

در این مرحله پس از انتخاب ماتریس جانشینی مناسب، از یک الگوریتم ژنتیکی برای تطبیق چندگانه استفاده می‌شود. در ادامه تعریف این الگوریتم ژنتیکی که شامل مشخص نمودن نمایش جمعیت، انتخاب جمعیت اولیه، تابع امتیازدهی و عملگرهای ترکیب و جهش می‌باشد، بیان می‌گردد.

- **نمایش جمعیت:** جمعیت مورد نظر در این الگوریتم همانند بسیاری از روش‌های دیگر با استفاده از ماتریس‌هایی مدل می‌شود که دارای n (تعداد کل رشته‌ها) سطر هستند و در هر سطر یک رشته‌ی تطبیق شده قرار خواهد گرفت. تعداد ستون‌های این ماتریس، طبق مشاهدات تجربی و همچنین کاربرد در سایر الگوریتم‌های ژنتیکی، حداکثر برابر ۱.۲ برابر طول بزرگترین دنباله‌ی مورد تطبیق قرار داده شده است.
- **انتخاب جمعیت اولیه و تابع امتیاز دهی:** در این کار برای اینکه هدف ما همگرایی سریع و استفاده از عملگرهای ساده برای افزایش سرعت الگوریتم ژنتیکی است، انتخاب یک جمعیت اولیه مناسب اهمیت پیدا می‌کند. در اینجا ابتدا تمام تطبیق‌های دوگانه‌ی هر یک از رشته‌ها را با رشته‌های دیگر با استفاده از ماتریس جانشینی انتخاب شده و الگوریتم ClustalW تشکیل می‌دهیم. استفاده از ماتریس جانشینی مناسب که نشان دهنده‌ی اطلاعات سراسری رشته‌هاست کمک می‌کند که این تطبیق‌های دوگانه نه به صورت محلی بلکه با استفاده از اطلاعات سراسری تشکیل شوند. همچنین استفاده از الگوریتم ClustalW که برخی از واقعیات زیستی اسیدهای آمینه را در نظر می‌گیرد کمک می‌کند که فواصل وارد شده در رشته‌های اولیه از نظر زیستی معقول باشند و در پاسخ نهایی تأثیر بد نداشته باشند. برای مثال تعداد هر ناحیه‌ی اسید آمینه بین دو ناحیه‌ی فاصله به سختی می‌تواند کمتر از ۸ اسید آمینه باشد، که این نکته در الگوریتم ClustalW مدنظر قرار داده شده است. در نهایت امتیاز هر یک از اعضاء جمعیت بر اساس مجموع وزن‌دار امتیاز جفت کاراکترهای منطبق شده با یکدیگر استفاده می‌شود و البته به ازای فواصل نیز جریمه‌ای مد نظر قرار داده می‌شود. جریمه‌ی فاصله مقداری است که از مجموع وزن دار امتیاز جفتی اسید آمینه‌ها کم شده است و برابر با یک جریمه‌ی اولیه (g) منهای طول فاصله (x) ضربدر جریمه‌ی گسترش فاصله (r) است. وزن (w_{ij}) در اینجا برابر ۱ برای تمامی رشته‌ها در نظر گرفته شده است ولی در نمونه‌هایی مانند [۴] میزان مشخصی دارد.

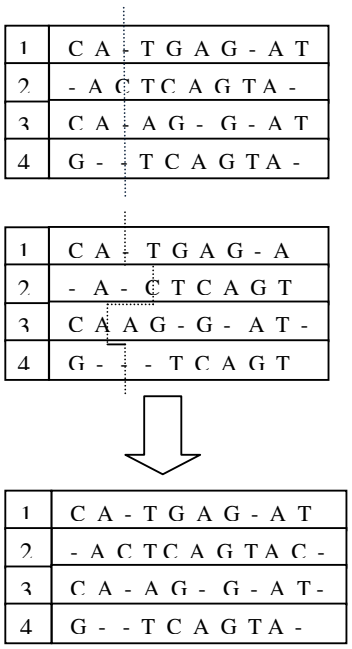
$$score = \sum_{i=2}^{k} \sum_{j=1}^{i-1} w_{ij} s_{ij} - (g + rx)$$

تابع امتیازدهی در هر الگوریتم ژنتیکی باید برای مشخص کردن میزان مناسب بودن اعضای جمعیت مشخص شود. در این مورد نیز میزان مناسب بودن یک عضو جمعیت بعنوان یک تطبیق چندگانه، امتیازی است که اسید آمینه‌های تطبیق داده شده بدست می‌آورند. در بسیاری از روش‌ها جریمه‌ای نیز برای فواصل در نظر گرفته شده است که در این مقاله هم از این جریمه استفاده می‌کنیم. استفاده از جریمه‌ی فاصله باعث می‌شود عملیات ژنتیکی به منظور بهبود امتیاز تطبیق، فواصل را بیش از اندازه به رشته‌ها وارد نکند. مخصوصاً در تعداد نسل‌های زیاد وجود چنین جریمه-ای لازم است تا پاسخ نهایی منحرف نشود.

در روش [۱] به این دلیل که تکیه‌ی اصلی روند بهبود الگوریتم بر عملیات جهش و وارد کردن فواصل است، اضافه کردن جریمه‌ی فاصله منجر به کمتر شدن امتیاز ژن‌هایی است که تغییر یافته‌اند و همین امر سبب شده تا در این روش از جریمه‌ی فاصله استفاده نشود. عدم استفاده از

جریمه‌ی فاصله سبب بوجود آمدن فواصل بیش از حد در جمعیت تکامل یافته و ماندن در بهینه‌ی محلی می‌شود. در اینجا این مشکل با تعریف تابع امتیازدهی مناسب و همچنین عملیات جهش مناسب – که در ادامه به آن اشاره می‌کنیم – رفع شده است.

- **عملیات ترکیب:** از دو نوع عملیات مرسوم برای تولید نسل بعدی در اینجا استفاده شده است. یکی عملیات ترکیب افقی و دیگری عملیات ترکیب عمودی. در ترکیب افقی دو والد انتخاب می‌شوند و به ازای هر رشته در فرزند، رشته‌ی متناظر از یکی از دو والد به صورت تصادفی انتخاب شده و جایگزین می‌گردد. در عملیات ترکیب عمودی، یک نقطه‌ی جهش انتخاب می‌شود. این نقطه والد اول را به دو قسمت تقسیم می‌کند که آنها را A و B می‌نامیم. به صورت مشابه در والد دوم دو قسمت 'A و 'B به دست می‌آید. برای اینکه بتوانیم نیمه‌ی A را با نیمه‌ی 'B ترکیب کنیم، باید سازگار بودن تعداد اسید آمینه‌ها در این دو نیمه را بررسی کنیم. اگر این سازگاری وجود نداشت، در 'B از نقطه‌ی انتخاب شده آنقدر جابجا می‌شویم که این سازگاری حاصل شود. در شکل ۳ این عملگر ترکیب نمایش داده شده است.



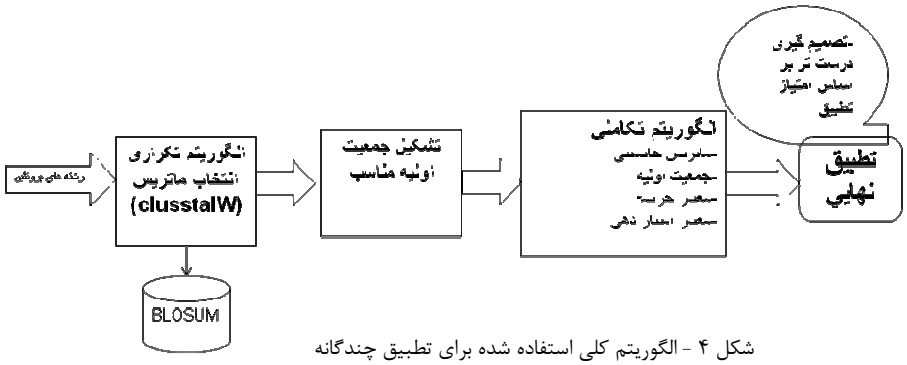
شکل ۳- عملیات ترکیب عمودی تعریف شده

این عملیات نسبت به عملیات ترکیب عمودی که در [۱] تعریف شده است از نظر اجرایی سریعتر می‌باشد. در [۱] اگر سازگاری وجود نداشت، الگوریتم به دنبال یک نقطه‌ی دیگر می‌گردد و این عمل را آنقدر تکرار می‌کند تا نقطه‌ی مناسب پیدا شود (در بدترین حالت این نقطه در ابتدای رشته‌ها انتخاب می‌گردد). در نتیجه از نظر اجرایی زمان این عمل در بدترین حالت از مرتبه‌ی (تعداد رشته‌ها × تعداد ستون‌های تطبیق به توان ۲) است؛ درحالیکه در روش معرفی شده این کنترل یک مرتبه و در زمان (تعداد رشته‌ها × تعداد ستون‌های تطبیق) صورت می‌گیرد و ترکیب انجام می‌پذیرد. بررسی تجربی نشان می‌دهد کارایی این دو عمل ترکیب یکسان است. مشکل دیگری که ترکیب معرفی شده در [۱] دارد این است که به دلیل اینکه ترکیب را تنها در صورت وجود سازگاری بین ژن‌های والد انجام می‌دهد، نقاطی که عملاً ترکیب عمودی در آنها رخ می‌دهد را به سمت نقاط انتهایی ژن‌ها متمایل می‌سازد؛ زیرا تعداد اسید آمینه‌ی مورد بررسی در این نقاط کمتر و در نتیجه احتمال سازگاری بیشتر است. این مسئله به صورت عملی نیز آزمایش شده است. این مشکل نیز در عملیات ترکیب جدید، با انجام قطعی ترکیب در هر نقطه‌ی انتخاب شده، رفع شده است.

- **عملیات جهش:** در عملیات جهش یک محل به طور تصادفی انتخاب شده و یک کاراکتر فاصله در آن محل داخل می‌شود. در واقع عملیات جهش عبارت است از باز کردن یک فاصله جدید، بستن فاصله موجود، گسترش فاصله موجود، و یا کم کردن سایز فاصله موجود. تغییری که در این قسمت از الگوریتم داده شده است این است که در تعیین جریمه‌ی فاصله در عملیات جهش، اولاً واقعیات زیستی در مورد محل قرار گرفتن فاصله‌ها مدنظر قرار داده شده است و ثانیاً سعی شده است به نقاطی که از نظر عمودی تطبیق مناسبی در آنها انجام شده است (مثلاً ستون‌های کاملاً منطبق شده) با احتمال بیشتری تغییر انجام نشود. در واقع قبل از اینکه یک رشته‌ی مشخص شده در معرض جهش قرار گیرد، احتمال رخداد جهش در نقاط مختلف آن محاسبه شده و به این ترتیب جهش انجام شده در مکان‌های محتمل واقعی اعمال می‌گردد. مزیت این عملیات سادگی آن است که باعث می‌شود بر خلاف عملگرهای پیچیده زمان کمتری در هر بار اجرا به خود اختصاص دهد.

روش کلی را می‌توان به طور خلاصه در شکل ۴ مشاهده کرد. دقت بیشتر در ایجاد جمعیت اولیه و همچنین عملگرها باعث می‌شود که امتیازی که برای بهبود الگوریتم ژنتیکی تعریف شده است، منجر به بهبود تطبیق از نظر زیستی نیز بشود.

در واقع در حالت کلی بهینه سازی چنین تابعی از نظر ریاضی ممکن است پاسخ مناسبی را از نظر معیارهای زیستی به دست ندهد ولی با اصلاح عملگرها، جمعیت اولیه و تابع امتیاز دهی در بخش نتایج خواهیم دید که بهبود تابع امتیازدهی الگوریتم ژنتیکی در اکثر موارد منجر به بهبود تطبیق از نظر زیستی نیز می‌شود.



شکل ۴ - الگوریتم کلی استفاده شده برای تطبیق چندگانه

۶ نتایج آزمایشات

تطبیق‌های انجام شده بوسیله‌ی الگوریتم ارائه شده با نتایج روش [۱] و همچنین ClustalW که استانداردی برای سنجش روش‌های تطبیق چندگانه است، مقایسه شده است.

در مقایسه از چهار مرجع از مجموعه داده‌ی استاندارد BAliBASE استفاده شده است. این رشته‌ها از مرجع ۱ از این مجموعه داده انتخاب شده‌اند که رشته‌هایی با کمتر از ۲۵٪ شباهت می‌باشند. دو گروه اول رشته‌هایی با طول کوتاه (حدود ۱۰۰)، گروه دوم رشته‌هایی با طول متوسط و آخرین گروه رشته‌هایی با طول بلند (حدود ۷۰۰) می‌باشند.

برای سنجش میزان مناسب بودن یک تطبیق، در این مجموعه داده، دو معیار SPS (درصد جفت اسید آمینه‌هایی که در تطبیق به دست آمده مطابق تطبیق مرجع هستند) و CS (درصد ستون‌هایی در تطبیق که کاملاً با ستون متناظر در مرجع یکسان هستند) در نظر گرفته شده است که که در جدول ۱ نتایج این معیارهای در دو سطر جداگانه برای هر مجموعه تست نشان داده شده است. این معیارها توسط برنامه‌ی bali_score.c که در این مجموعه داده قرار داده شده است اندازه گیری شده‌اند.

نتایج حاصل از متوسط گیری بین ۵ اجرای برنامه می‌باشند. در هر اجرا ۲۰۰۰ بار ارزیابی تابع امتیازدهی انجام شده است. در دو ستون اول جواب بهتر پر رنگ شده است، و بین سه ستون دوم که یکی پاسخ clustalW است و دو تای دیگر از این پاسخ به عنوان یک پاسخ اولیه استفاده می‌کنند نیز پاسخ بهتر پررنگ شده است.

برای انجام این آزمایشات احتمال وقوع عملیات ترکیب و جهش به صورت تجربی مقدار دهی شده است.

همانطور که ملاحظه می‌شود نتایج الگوریتم جدید از الگوریتم [۱] در حالتی که از هسته‌ی ClustalW استفاده نشود معمولاً بهتر بوده است. همچنین در استفاده از هسته‌ی اولیه، معمولاً پس از اجرای الگوریتم نتیجه نسبت به هسته‌ی اولیه بهبود بیشتری نسبت به الگوریتم [۱] یافته است. این در حالی است که الگوریتم [۱] در برخی موارد کیفیت هسته‌ی اولیه را تقلیل داده است. دلیل این امر این است که در تابع امتیازدهی تعریف شده در [۱] جریمه‌ای برای فواصل در نظر گرفته نشده است و وارد شدن کاراکترهای فاصله در هسته‌ی اولیه بدون در نظر گرفتن معیارهای دقیق منجر به کاهش کیفیت آن شده است.

همین امر نشان می‌دهد در الگوریتم ارائه شده، روند تکاملی الگوریتم و نحوه‌ی بهینه سازی تناسب بیشتری با پاسخ صحیح داشته است و همین مسئله اطمینان به آن برای اجراهای بعدی را بیشتر می‌سازد.

ذکر این نکته مهم است که با بیشتر کردن تعداد نسل‌های الگوریتم ژنتیکی نتایج بخش‌های بدون هسته‌ی اولیه نیز بسیار بهبود پیدا می‌کند اما در این کار به این دلیل که هدف یافتن یک پاسخ مناسب در زمان کم (۲۰۰۰ نسل) می‌باشد نتایج به این مقادیر محدود شده‌اند.

	الگوریتم [۱]	الگوریتم جدید	الگوریتم [۱] با هسته ClustalW	الگوریتم جدید با هسته ClustalW	ClustalW
1aab (sps)	0.617	**0.717**	0.852	**0.854**	0.818
1aab (cs)	0.421	**0.553**	0.764	**0.768**	0.673
1tvxA (sps)	0.213	**0.237**	0.228	**0.239**	0.220
1tvxA (cs)	**0.025**	0	**0.027**	0	0
kinase (sps)	0.284	**0.299**	0.628	0.645	**0.655**
kinase (cs)	**0.119**	0.091	0.473	**0.485**	**0.485**
1ped (sps)	0.414	**0.542**	0.559	**0.763**	0.748
1ped (cs)	0.22	**0.394**	0.418	**0.695**	0.666

جدول ۱ – نتایج آزمایشات بر روی چند زیر مجموعه از مجموعه داده‌ی Balibase2

۷ کارهای آینده

همانطور که گفته شد، نیاز به روش‌های سریع و کارا برای تطبیق چندگانه از مسائل مهم زیستی است که الگوریتم مناسبی برای انجام آن وجود ندارد. ارائه روشی سریع، که نیاز به وجود پیش فرض‌ها و اطلاعات قبلی برای حل مسئله نداشته باشد و بتواند پاسخی که مطابق با معیارهای زیستی است را به دست بدهد از اهمیت بسیاری برخوردار است. بالا بردن دقت تطبیق چندگانه نیز از مسائل مهم در این زمینه است.

معیارهای مختلفی می‌تواند به بهبود کیفیت تطبیق چندگانه در رشته‌های پروتئین بیفزاید. یکی از این موارد استفاده از اطلاعات ساختاری رشته‌های پروتئین برای تطبیق آنهاست.

در واقع رشته‌های پروتئین از لحاظ ساختاری خصوصیاتی دارند که ساختار دوم و سوم آنها نامیده می‌شود. گاهی اطلاعات ساختاری رشته‌ها در دست است و می‌توان از این اطلاعات برای تطبیق چندگانه‌ی آنها استفاده کرد، مانند کاری که در [۱۱] و [۱۲] انجام شده است. توجه به اطلاعات ساختاری می‌تواند زمینه‌ای برای ادامه‌ی این کار باشد.

یکی دیگر از مسائلی که برای بهبود دقت در تطبیق چندگانه مطرح است استفاده از اطلاعات عمودی در نحوه‌ی محاسبه‌ی امتیاز تطبیق-هاست. منظور از اطلاعات عمودی، توجه به اسید آمینه‌های منطبق شده در کل رشته‌ها و همچنین توجه به نواحی است که میزان انطباق در آنها بیشتر است. در [۵] تابعی برای امتیاز دهی تطبیق‌ها معرفی شده است که از این اطلاعات استفاده می‌کند. بعنوان کار آینده برای این کار در نظر داریم این تابع امتیازدهی را در الگورتیم مد نظر قرار دهیم تا اثر چنین معیارهایی را بر تطابق پاسخ با معیارهای زیستی بررسی کنیم.

تقدیر

در اینجا لازم می‌دانیم از آقایان سجاد شیرعلی شهرضا، دانشجوی دکترای دانشکده‌ی مهندسی کامپیوتر دانشگاه صنعتی شریف، و هادی محضرنیا، دانشجوی کارشناسی ارشد مهندسی کامپیوتر دانشگاه صنعتی امیرکبیر، که در مرور و اصلاح این مقاله کمک بسیاری نمودند و همچنین در طول انجام این پروژه ما را راهنمایی کردند تشکر نماییم.

مراجع

[1] C. Gondro, B.P. Kinghorn; "**A simple genetic algorithm for multiple sequence alignment**"; Genetics and Molecular Research 6(4): 964-982, (2007)

[2] Ren´e Thomsen, Gary B. Fogel, Thiemo Krink; "**A Clustal Alignment Improver using Evolutionary Algorithms**"; Proceeding of forth congress on evolutionary computation (ECE-2002) , vol 1. 121-126, (2002)

[3] L. A. Anbarasu, P. Narayanasamy† and V. Sundararajan; "**Multiple molecular sequence alignment by island parallel genetic algorithm**"; CURRENT SCIENCE, VOL. 78, NO. 7, 10, (2000)

[4] J D Thompson, D G Higgins, T J Gibson; "**CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**"; *Nucleic Acids Res.*; **22**:4673-4680, (1994)

[5] Yue Lu, Sing-Hoi Sze; "**Improving accuracy of multiple sequence alignment algorithms based on alignment of neighboring residues**"; Neucleic Acids Research, 1-10, (2008)

[6] Iain M. Wallace, Orla O'Sullivan, Desmond G. Higgins and Cedric Notredame; "**M-Coffee: combining multiple sequence alignment methods with T-Coffee**"; Nucleic Acids Research, Vol. 34, No. 6, (2006)

[7] Notredame, C., Holm, L. & Higgins, D. G.; "**consistency based objective function for alignment evaluation**"; Bioinformatics, 14, 407-422, (1998)

[8]C. Notredame, D. Higgins, J. Heringa ; "**T-Coffee: A novel method for multiple sequence alignments**"; Journal of Molecular Biology, 302, 205-217, (2000)

[9] Cédric Notredame* and Desmond G. Higgins; **" SAGA: sequence alignment by genetic algorithm";** Nucleic Acids Research, Vol. 24, No. 8, (1996)

[10] Thompson, J. D., Koehl, P., Ripp, R. and Poch, O.; "**BAliBASE3.0: latest developments of the multiple sequence alignment benchmark**"; Proteins vol. 61; 127-136 (2005)

[11] Litvinov, Mironov A.A, Finkelstein A.V., Roytberg M.A.; "**Information about secondary structure improves quality of protein alignment** "; Journal of Molecular Biology (2006)

[12] HUZEFA RANGWALA and GEORGE KARYPIS; "**fRMSDAlign: Protein Sequence Alignment Using Predicted Local Structure Information for Pairs with Low Sequence Identity**"; WSPC – Proceedings: 22-25, (2007)

[13]Higgins,D.G. Bleasby,A.J. and Fuchs,R.; "**CLUSTAL V: improved software for multiple sequence alignment**"; CABIOS, vol .8, 189-191, (1991)

[14] F. CORPET; "**Multiple sequence alignment with hierarchical clustering**"; Nucl. Acids Res., 16 (22), 10881-10890, (1988)

[15] V. A. Simossis1 and J. Heringa; "**PRALINE: a multiple sequence alignment toolbox that integrates homology-**extended and secondary structure information"; Nucleic Acids Res > v.33(Web Server issue); (2005)

[16] Brocchieri L, Karlin S; "**Asymetric-iterated multiple alignment of protein sequences**"; JMB 276, 249-264 (1998).

[17] ftp://ftp.ncbi.nih.gov/blast/matrices

[18] Peter Clote, Rolf Backofen; "**Computational Molecular Biology**"; John Wiley Press, 2000

[19] Needleman SB, Wunsch CD; "**A general method applicable to the search for the similarities in the amino acid sequence of two proteins**"; Journal of Mol. Biol. 48. 443-453, (1970).