مجله دانش و تندرستی در علوم پایه پزشکی، دوره ۱۶، شماره ۱، بهار ۱۴۰۰
دانشگاه علوم پزشکی و خدمات بهداشتی درمانی شاهرود

DOI: 10.22100/jkh.v16i1.2512

مقاله پژوهشی

ارایه یک روش جدید انطباق چندگانه توالی‌های دی‌ان‌ای و پروتئین براساس الگوریتم‌های تکاملی

نازنوش‌سادات اطمینان[۱]، الهام پروین‌نیا*[۱]، علی شریفی‌زارچی[۲]

۱- گروه مهندسی کامپیوتر- واحد شیراز- دانشگاه آزاد اسلامی- شیراز- ایران.

۲- گروه مهندسی کامپیوتر- دانشگاه صنعتی شریف- تهران- ایران.

تاریخ دریافت: ۱۳۹۹/۱۲/۱، **تاریخ پذیرش:** ۱۴۰۰/۱/۱۶

چکیده

مقدمه: *مطالعه حیات و آشکارسازی وظایف ژن‌ها یک مسأله مهم در محاسبات زیستی است. در انطباق توالی‌های زیستی، برای شناسایی ژن‌ها، اندازه‌گیری شباهت بین توالی‌ها انجام می‌شود. وقتی با مسأله اندازه ژنوم در انطباق‌های چندگانه مواجه می‌شویم، با مشکل کمبود حافظه و افزایش زمان روبه‌رو هستیم. بنابراین، روشی که بتواند سریع و بدون کاهش دقت، انطباق ژنوم‌ها را داشته باشد، تأثیر به‌سزایی در تحلیل توالی‌ها خصوصاً توالی‌های بلند را همراه دارد.*

مواد و روش‌ها: *ابتدا روشی را برای تقسیم هر توالی به زیر توالی‌های کوتاه معرفی می‌کنیم. سپس از الگوریتم‌های تکاملی برای انطباق زیرتوالی‌ها استفاده می‌کنیم.*

نتایج: *روش پیشنهادی در هفت مجموعه داده با تعداد نکلوئتیدهای مختلف به‌ازای هر توالی دی‌ان‌ای و افزایش تدریجی از ۱۸۰۰۰ تا ۱۴ میلیون نکلئوتید، ارزیابی شده و با پنج روش مشهور انطباق چندگانه مقایسه شده است. بالاترین میزان دقت برای باکتری variola با میزان ۰/۹۳ و بالاترین سرعت انطباق ۰/۶ بر حسب دقیقه برای این باکتری است.*

نتیجه‌گیری: *اکثر روش‌های انطباق چندگانه در توالی‌های کوتاه یا تعداد کم، دقت مناسبی دارند اما برای دنباله‌های طولانی‌تر به قدرت محاسباتی بالایی نیاز دارند. الگوریتم پیشنهادی با انطباق توالی‌های بلند، در زمانی قابل قبول و حفظ دقت و همچنین استفاده بهینه از حافظه، بر این نقص غلبه می‌کند.*

واژه‌های کلیدی: انطباق چندگانه توالی، داده ژنوم کامل، تقسیم توالی، الگوریتم‌های تکاملی.

***نویسنده مسئول**: شیراز، شهر صدرا، پردیس دانشگاه آزاد اسلامی واحد شیراز، دانشکده مهندسی۱، گروه کامپیوتر، **تلفن:** ۰۷۱۳۶۴۱۰۰۴۱، **نمابر:** ۰۷۱۳۶۴۱۰۰۵۹، **Email:** eparvinnia@yahoo.com

ارجاع: اطمینان نازنوش‌سادات، پروین‌نیا الهام، شریفی‌زارچی علی. ارائه یک روش جدید انطباق چندگانه توالی‌های دی‌ان‌ای و پروتئین براساس الگوریتم‌های تکاملی. مجله دانش و تندرستی در علوم پایه پزشکی ۱۴۰۰؛۱۶(۱):۲۰-۱۳.

مقدمه

یکی از مسائل مهم در علوم زیست‌شناسی محاسباتی، بررسی و مطالعه ژن‌هاست، کشف ژن‌ها به شناسایی، تشخیص و حتی درمان بیماری‌ها کمک ویژه‌ای می‌کند. یکی از مسائل مهم در این حوزه انطباق توالی‌هاست که برای پیش‌بینی ساختار پروتئین، عملکرد مولکولی، واکنش‌های بین مولکولی و پیدا کردن ژن کاربرد دارد (۱). انطباق توالی یعنی مقایسه و یافتن همسانی بین کاراکترهای دو توالی که این کاراکترها نکلوئتید برای DNA و آمینو اسید برای پروتئین‌ها هستند. اگر این مقایسه روی دو توالی اتفاق بیافتد انطباق دوگانه و اگر بیشتر از دو توالی شود، انطباق چندگانه است (۲). مشکل اصلی انطباق چندگانه از نظر دانش کامپیوتر، تعداد زیاد مقایسه‌های موردنیاز هنگام جستجوی شباهت‌هاست. محققان به علت این افزایش شدید و رو به رشد داده‌های مولکولی به روش‌های محاسباتی با کارایی بالا روی آوردند (۳).

به طور کلی، سه دسته از الگوریتم بهینه‌سازی دقیق (Exact)، پیش‌رونده (Progressive) و تکرارشونده (Iterative) برای انطباق چندگانه مورد استفاده قرار می‌گیرد. روش‌های دقیق براساس برنامه‌ریزی پویا و ایجاد یک ماتریس امتیازدهی بهترین انطباق را ارائه می‌دهند، ولی با بزرگ شدن ابعاد داده‌ها از پیچیدگی محاسباتی بالایی برخوردار هستند. یکی از معروف‌ترین این روش‌ها اسمیت واترمن (۸) است. پراستفاده‌ترین روش برای انطباق چند توالی از یک جستجوی اکتشافی به نام روش پیشرونده (روش سلسله مراتبی یا درختی) استفاده می‌کند، که انطباق نهایی را از روی ترکیب انطباق‌های دو به دوهایی که با شبیه‌ترین جفت شروع می‌شوند و تا دورترین جفت‌ها ادامه پیدا می‌کند، می‌سازد که روش clustalw (۴) و Halign-II (۹)و muscle (۵) نمونه‌هایی از این گروه هستند. تمام انطباق‌های پیشرونده به دو مرحله احتیاج دارند: مرحله اول که در آن فاصله بین توالی‌ها به وسیله یک درخت که درخت راهنما (Guide tree) نام دارد، نمایش داده می‌شود و مرحله دوم که در آن انطباق نهایی با توجه به درخت راهنما و از اضافه کردن توالی‌ها به یکدیگر به دست می‌آید. انطباق‌های پیشرونده نمی‌توانند بهینه کلی باشند. مشکل اصلی این است که وقتی خطایی در هر یک از مراحل ساخت انطباق نهایی رخ می‌دهد، این خطا به مراحل نهایی انتشار پیدا می‌کنند (۳). مجموعه‌ای از روش‌های تولید انطباق چندگانه که خطاهای ناشی از الگوریتم‌های پیشرونده را کاهش می‌دهند در زمره روش‌های تکرار شونده قرار می‌گیرند، زیرا عملکردشان بسیار شبیه به روش‌های پیشرونده است با این تفاوت که مرتباً توالی‌های اولیه را دوباره انطباق می‌دهند و به انطباق چندگانه اضافه می‌کنند که روش‌های Dalign (۱۰) و Mafft (۶) از معروف‌ترین این روش‌ها هستند. بعضی از روش‌ها مانند روش Fame (۷) با استفاده تلفیقی از روش‌های تکرار شونده و شکستن توالی‌ها براساس اصل تقسیم و غلبه سعی در حل مسأله در ابعاد کوچک‌تر نموده‌اند. روش‌های دیگری هم برای حل مسأله انطباق وجود دارند که از الگوریتم‌های فراابتکاری استفاده می‌کنند مانند روش حل مسئله انطباق چندگانه چند هدفه مبتنی بر الگوریتم زنبور عسل (۱۱)، روش حل مسأله انطباق چندگانه با بهره‌گیری از الگوریتم بهینه‌سازی ازدحام ذرات (۱۲) و روش بهینه‌سازی ازدحام ماهی (۱۳) که براساس دو تابع شباهت برای حفظ کیفیت انطباق استفاده می‌کند که شامل مجموع وزن جفت توالی‌ها برای مناطق مشابه افقی و عمودی است. استفاده از یک روش تکه تکه شده توالی دو مرحله‌ای با الگوریتم ازدحام ذرات (۱۴) که با انجام داده انطباق در دو مرحله و ارسال خروجی مرحله اول به عنوان ورودی مرحله دوم سبب بالارفتن کیفیت انطباق شده است. روش ViralMSA (۱۸) ابزاری کاربر پسند با راهنمای مرجع است که از تکنیک‌های الگوریتمی نقشه بردارها برای فعال کردن مجموعه داده‌های ژنوم ویروسی بسیار بزرگ استفاده می‌کند. با تعداد دنباله‌ها به صورت خطی مقیاس بندی می‌شود و قادر است ده‌ها هزار ژنوم کامل ویروسی را در زمان کوتاهی انطباق دهد. روش CONSENT (۱۹) روشی است که به طور مؤثر در Readهای بسیار طولانی مقیاس بندی می‌شود و به شما اجازه می‌دهد در طی ۱۰ روز یک مجموعه داده کامل انسانی را پردازش کنید، با این حال این روش دارای نرخ خطا می‌باشد. روش BL-ABC (۲۰) یک الگوریتم دو سطحی زنبورعسل مصنوعی می‌باشد. که براساس آموزش مدل مارکوف پنهان بر روی داده‌های پروتئینی می‌باشد. مدل تصادفی آموزش دیده ایجاد شده، ماتریس‌های احتمال وابسته به موقعیت را در نسبت‌های بالاتر پیش بینی می‌کند. روش EBSODP (۲۱) برای محاسبه بهینه انطباق توالی‌ها ساخته شده است، اما به دست آوردن دقت مطلوب هنوز هم چالش انطباق چندگانه است. برای رفع این کمبود در همگرایی زودرس، این روش یک اندازه جدید جمعیت پویا را ارایه می دهد. این مکانیسم پیشرفته به صورت پویا راه حل تعیین شده در فضای جستجو برای هر تکرار را برای حفظ تنوع جمعیت کاهش می‌دهد.

در حالی که تمام روش‌های ذکر شده مزایا و معایب خود را دارند، اما نمی‌توانند با توالی‌های با طول بلند کار کنند و در مواجهه با ژنوم کامل پیچیدگی محاسباتی بالایی دارند. هدف کلی این مقاله انطباق توالی‌های متعدد با طول‌های طولانی با حفظ دقت و بهبود سرعت قابل قبول نسبت به الگوریتم‌های دیگر است. به طور خلاصه، بخش‌های اصلی این مقاله عبارت‌اند از: ۱- ارائه یک الگوریتم سریع برای انطباق چندین توالی طولانی با یافتن زیر دنباله‌های مشترک در کل مجموعه داده. ۲- ارایه مکانیزم تقسیم برای تقسیم مناسب توالی به زیر دنباله‌های کوچک‌تر ۳- استفاده از یک روش ترکیبی از الگوریتم‌های فراابتکاری برای انطباق زیر توالی‌ها. ۴- پیشنهاد یک تابع هدف که ترکیبی از بیشینه کردن کیفیت انطباق زوج توالی‌ها و پاداش دادن به شکاف‌های

کمتر در انطباق نهایی است. در ادامه مقاله، سایر بخش‌ها به شرح زیر است. در بخش ۲ روش پیشنهادی معرفی می‌شود و به دنبال آن در بخش ۳ به بحث روش می‌پردازیم و در ادامه در بخش ۴ به ارزیابی نتایج و در نهایت در بخش ۵ به نتیجه‌گیری پرداخته‌شده‌است.

مواد و روش‌ها

در این مقاله، ما یک تابع هدف (Objective function) را برای استفاده در مرحله بهینه‌سازی معرفی می‌کنیم، به عنوان یک جمع وزنی از دو شاخص طبیعی شده: ۱- نرمال شده، تابع مجموع زوج توالی‌ها (Sum of pair score (SP))که یک روش معمول برای بررسی میزان کیفیت انطباق توالی‌ها است ۲- (sp) تعداد ستون‌هایی که در آن‌ها همه سمبل‌ها یکسان و مشابه هستند. (Total score) تا انطباقی با حداقل شکاف‌های ممکن پیدا کنند. (st) تابع هدف و میزان امتیازات در روابط ۱، ۲ و ۳ مشاهده می‌شود.

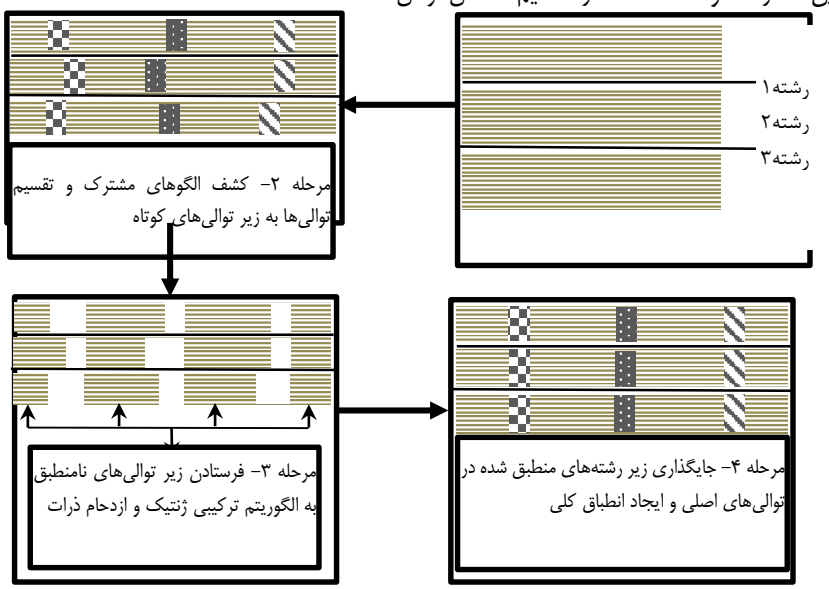
$$\mathbf{OF} = \mathbf{C_1} \times \frac{\mathbf{SP}}{\mathbf{SP_{Max}}} + \mathbf{C_2} \times \frac{\mathbf{ST}}{\mathbf{ST_{Max}}}$$

$$\mathbf{SP} = \sum_{k=1}^{L} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \mathbf{score}\ (s_{ik}, s_{jk})$$

$$\boldsymbol{ST} = \sum_{j=1}^{L} \boldsymbol{T_j}$$

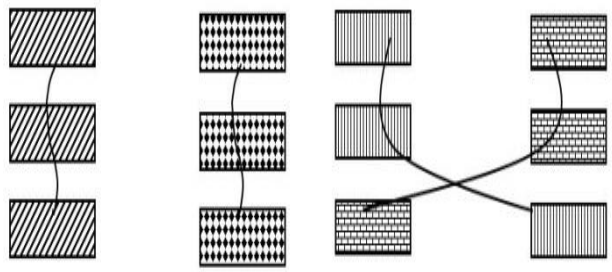
در رابطه ۱ مقدار $\frac{SP}{SP_{Max}}$ و $\frac{ST}{ST_{Max}}$ نرمال شده مقادیر SP و ST است. از آنجایی که هر دو مقدار SP و ST در مجموعه داده‌های مختلف می‌توانند مقادیر متفاوتی داشته باشند، ما آن‌ها را نرمال‌سازی کردیم. میزان SPmax ، STmax بهترین مقدار SP و ST هستند. از تقسیم حاصل نرمال شده مقادیر آن‌ها به دست می‌آید. در این تابع هدف، وزن‌های C1 و C2 به ترتیب مقادیر ۰/۸ و ۰/۲ را در نظر گرفتیم، تا تأثیرات آن‌ها بر روی تابع هدف مشخص شود. این مقادیر را تجربی از بین تأثیر مقادیر در بازه (۰ و ۱) به دست آورده‌ایم. در رابطه ۲، s_{ik} و s_{jk} به ترتیب سمبل‌ها در توالی Si و Sj می‌باشند. Score میزان امتیازی است برای مقایسه مقادیر جفت دنباله‌ها که میزان امتیاز ۲ را بر اساس تطابق، و میزان امتیاز ۱ را برای عدم تطابق و میزان امتیاز ۰ را برای شکاف‌ها اختصاص می‌دهد. در رابطه ۳، مقدار ST برای محاسبه تعداد ستون‌هایی است که در هر ستون دارای سمبل‌های یکسان هستند که توسط T_j محاسبه می‌شود. بدین‌صورت که به هر ستون مشابه امتیاز ۱ و به سایر ستون‌ها امتیاز ۰ تعلق می‌گیرد. علاوه بر این، L و m به ترتیب طول و تعداد توالی را نشان می‌دهند. هدف از حل این مسئله به حداکثر رساندن تابع هدف (رابطه ۱) است.

ما یک روش تقسیم توالی را به‌عنوان گام اولیه در مسأله انطباق چندگانه برای توالی‌های ژنوم پیشنهاد می‌کنیم، در نتیجه، طول توالی‌ها محدودیتی در روش پیشنهادی ما محسوب نمی‌شوند. الگوریتم پیشنهادی در شکل ۱ ارائه شده است و شامل چهار مرحله است: مرحله ۱- زیر دنباله‌های مشترک از همه توالی‌ها استخراج می‌شوند. مرحله ۲- توالی‌ها با توجه به الگوهای مشترک استخراج شده به چند قسمت تقسیم می‌شوند. مرحله ۳- ترکیب الگوریتم ژنتیک و ازدحام ذرات که در زیر بخش‌های بعدی توضیح داده شده است، برای انطباق زیر توالی‌های کوتاه شده در هر قسمت اعمال می‌شود. مرحله ۴- پس از انطباق تمام زیرتوالی‌ها، آن‌ها ادغام می‌شوند تا انطباق نهایی را تشکیل دهند.



شکل ۱- الگوریتم پیشنهادی

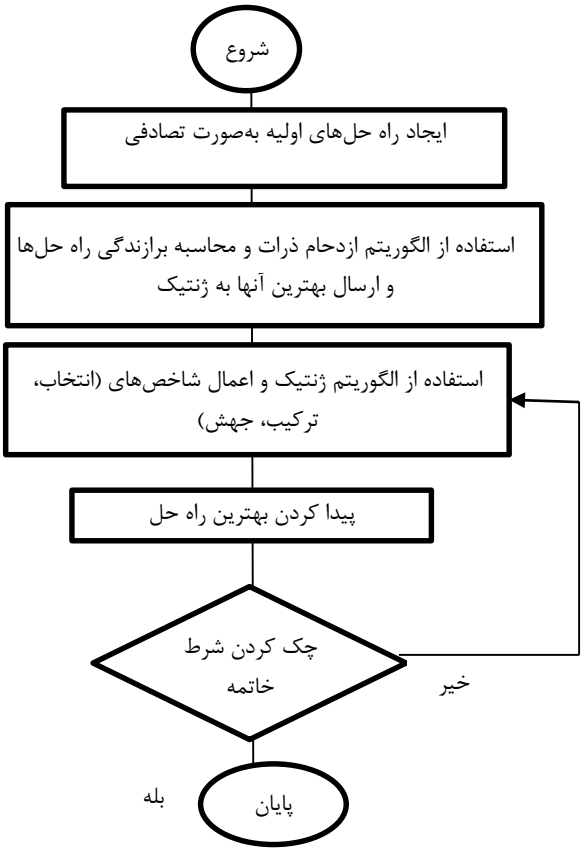
هدف ما یافتن بهترین مکان در توالی‌ها، با یافتن زیر دنباله‌های کوتاه است که بیشترین شباهت را در تمام توالی‌ها دارند. از این رو ما از مفهومی به نام الگو استفاده می‌کنیم. پیدا کردن مکان‌های صحیح شکستن هر توالی، بخش مهمی از تقسیم عمودی توالی‌ها است. استفاده از الگو یک راه ساده و قدرتمند برای انجام این کار است. الگو یک زیر رشته کوتاه است که در یک توالی قرار دارد و از یک الگوی خاص پیروی می‌کند. الگو نمایشگر یک رشته کوتاه از ۱ هاست که به‌عنوان یک ماسک عمل می‌کند. که سمبل‌های رشته اصلی براساس این ماسک انتخاب می‌شوند. تعداد ۱ها در الگو را وزن الگو می‌نامیم. برای یک توالی تمام زیر رشته‌های آن را با حرکت دادن الگو از ابتدا تا انتها و استخراج الگوهای متناظر هر محل به‌دست آوریم. تمام الگوهای روی یک توالی برای ما مهم نیستند بلکه الگو ای که در تمام توالی‌ها رخ می‌دهند برای ما مهم هستند که به آن‌ها الگوهای مشترک می‌گوییم. پس از یافتن الگوهای مشترک در همه توالی‌ها، مسیرها را از توالی اول تا توالی آخر ایجاد می‌کنیم، به شرط آنکه با مسیرهای دیگر برخورد نکنند (شکل ۲ ب). پس از استخراج الگوهای مشترک، توالی‌ها به‌صورت عمودی به چند قسمت تقسیم می‌شوند. سپس، هر قسمت با استفاده از الگوریتم بهینه‌سازی منطبق می‌شود.



شکل ۲ الف- مسیرهای مشترک دارای برخورد، شکل ۲ ب- مسیرهای مشترک بدون برخورد

در این مقاله از ترکیب الگوریتم ازدحام ذرات (۱۵) به‌عنوان یک روش استفاده از هوش جمعی و همچنین الگوریتم ژنتیک (۱۶) به‌عنوان یک الگوریتم تکاملی استفاده‌شده است که در شکل ۳ نشان داده‌شده است.

روش کار بدین صورت است که ابتدا راه‌حل‌های اولیه به‌صورت تصادفی ایجاد می‌شوند. سپس این داده‌های اولیه به‌عنوان ورودی وارد الگوریتم ازدحام ذرات شده و تابع برازندگی هر راه‌حل محاسبه می‌شود. بهترین‌های این مرحله، که مجموعه‌ای از بهترین‌های سراسری (Global best) می‌باشند، به الگوریتم ژنتیک داده می‌شوند و توابع انتخاب، ترکیب و جهش روی آن‌ها اعمال می‌شود. با این عمل ترکیبی دو مرحله‌ای ورودی‌های خام یک بار به شرایط مطلوبی رسیده‌اند و در گام بعدی با اعمال شاخص‌های جهش و ترکیب در ژنتیک میزان کیفیت انطباق بسیار مطلوب‌تر می‌شود. در انتها بهترین راه‌حل انتخاب می‌شود. همان‌طور که در پیشینه بیان شد، در واقع تضمین روش fame در گرو کارآمدی ابزاری بود که برای انطباق، انتخاب می شد. از این رو به‌کارگیری روشی ترکیبی از روش‌های فراابتکاری توانست به این نقص غلبه نماید.



شکل ۳- الگوریتم ترکیبی ازدحام ذرات و ژنتیک

بحث

این بخش صحت روش پیشنهادی را برای مسئله انطباق چندگانه توالی‌ها ارزیابی می‌کند. تمام نتایج توسط سیستمی با پردازنده Intel x64 Core i5 با سرعت کلاک ۲/۴ گیگاهرتز و ۸ گیگابایت حافظه RAM تحت سیستم‌عامل لینوکس اوبونتو به‌دست آمده است.

مجموعه داده‌های زیر برای آزمایش دقت و سرعت روش پیشنهادی در مقایسه با سایر روش‌ها در نظر گرفته شده است. دلیل انتخاب ما برای این ۷ مجموعه دارابودن ژن کامل و افزایش تدریجی سایز طول توالی‌هاست به‌طوری که باکتری variola با کمترین تعداد طول رشته در حدود ۱۸۰۰۰۰ و sorangium با بالاترین طول رشته در حدود ۱۴ میلیون نکلئوتید، به ازای هر یک رشته می‌باشند. تمامی مجموعه داده از سایت taxonomy browser قابل دریافت و ذخیره می‌باشند. ویژگی هرکدام از داده‌ها در جدول ۱ آمده است.

جدول ۱- مجموعه داده‌های مورد استفاده

مرجع	میانگین طول توالی	تعداد توالی	نام کامل ژن‌ها به همراه جزئیات باکتری‌های انتخابی
Taxonomy browser	۱۸۰۰۰۰	۴	DQ437585.1 Variola virus strain India 1964 7124 Vellore DQ441428.1 Variola virus strain India 1953 (New Delhi) DQ437586.1 Variola virus strain India 1964 7125 Vellore DQ441427.1 Variola virus strain India 1953 (Kali-Muthu-M50 Madras)
Taxonomy browser	۱۰۰۰۰۰۰	۴	NC_018495.1 Mycoplasma genitalium M2321 NC_018498.1 Mycoplasma genitalium M2288 NC_018497.1 Mycoplasma genitalium M6320 NC_018496.1 Mycoplasma genitalium M6282
Taxonomy browser	۲۰۰۰۰۰۰	۴	NC_003028.3 Streptococcus pneumoniae TIGR4 NC_012468.1 Streptococcus pneumoniae 70585 NC_012468.1 Streptococcus pneumoniae 70585 NC_012467.1 Streptococcus pneumoniae P1031
Taxonomy browser	۴۵۰۰۰۰۰	۴	NC_012759.1 Escherichia coli BW2952 NC_000913.2 Escherichia coli str. K-12 substr. MG1655 NC_010473.1 Escherichia coli str. K12 substr. DH10B NC_012967.1 Escherichia coli B str. REL606
Taxonomy browser	۱۴۰۰۰۰۰۰	۴	NZ_CP012672.1 Sorangium cellulosum strain So ce836 chromosome NZ_CP012673.1 Sorangium cellulosum strain So ce26 chromosome NZ_CP012670.1 Sorangium cellulosum strain So ceGT47 chromosome NC_010162.1 Sorangium cellulosum So ce56
(۱۷)	۲۰۰۰۰۰۰	۴	NC_003112.2 Neisseria meningitidis MC58 chromosome NC_010120.1 Neisseria meningitidis 053442 NC_008767.1 Neisseria meningitidis serogroup C FAM18 NC_013016.1 Neisseria meningitidis alpha14
GenBank	۳۳۰۰۰	۵۰	XP_0144230… PREDICTED: titin isoform X1.. X50

تابع هدف دو پارامتر وزن‌دهی C1 و C2 جهت طبیعی‌سازی دارد که C1 + C2 = 1 است. آن‌ها را براساس آزمایش تجربی از بین مقادیر (۰/۱ ۰/۲ ۰/۱) به‌دست آورده‌ایم که با آزمایش روی باکتری E.coli بهترین مقادیر وزن‌های C1 و C2 را به‌ترتیب ۰/۸ و ۰/۲ در نظر گرفتیم.

همچنین، وزن الگوهای مختلف روی کارایی روش پیشنهادی تأثیر دارد (۷). وقتی که وزن الگوها خیلی کوتاه است مقدار پوشش خیلی کم است به دلیل اینکه الگوهای مشترک توسط نویزها از بین می‌روند و با افزایش وزن الگو مقدار پوشش شروع به بهبود می‌کند و آزمایش تجربی به ما نشان داد برای وزن الگویی به طول ۱۱ ما نزدیک به ۷۶٪ پوشش می‌رسیم. با ادامه افزایش طول الگو مقدار پوشش به آهستگی شروع به کاهش می‌کند که این در اثر از دست دادن ناحیه‌های مشترکی که کوتاه‌تر از وزن الگو هستند، است که در شکل ۴ نشان داده شده است.

شاخص‌های دیگر، شاخص‌های مؤثر بر عملکرد الگوریتم ژنتیک است که شامل اندازه جمعیت، ترکیب (Cross over) و مکانیسم جهش (Mutation)، است. ما الگوریتم پیشنهادی با ژنتیک را روی باکتری E.coli با مقدار متفاوت شاخص‌های ذکر شده ارزیابی کردیم و بهترین مقادیر را براساس جدول ۲ به‌دست آوردیم.

شکل ٤- انتخاب وزن الگو

جدول ۲- شاخص‌های الگوریتم ژنتیک

داده ارزیابی	اندازه جمعیت	احتمال ترکیب	احتمال جهش	تابع هدف
E.coli	۱۰۰	۰/۷۵	۰/۲۵	۰/۸۹
	۲۰۰	۰/۷۵	۰/۲۵	۰/۹۱
	۳۰۰	۰/۷۵	۰/۲۵	۰/۹۳

الگوریتم ازدحام ذرات برای تشخیص موقعیت بعدی ذره، مجموع میزان موقعیت فعلی ذره همراه با ضریب w1، به اضافه موقعیت بهترین محلی به همراه ضریب w2 و موقعیت بهترین سراسری به همراه ضریب w3 را محاسبه می‌نماید. میزان این وزن‌ها در روش ازدحام جمعیت که میزان جستجوی سراسری (Exploration) و جستجوی محلی (Exploitation) را در الگوریتم کنترل می‌کنند، باید مقادیر مناسبی انتخاب شوند. با افزایش وزن w3 ما روی راه‌حل‌های سراسری تمرکز می‌کنیم و این باعث افزایش جستجو در فضای کلی مسأله می‌شود و با افزایش وزن w2 و تمرکز روی راه‌حل‌های محلی میزان جستجو در فضای محلی را افزایش می‌دهیم. بنابراین با تغییرات وزن‌ها می‌توانیم کنترل مناسبی روی فضای جستجوی سراسری و محلی در مسئله داشته باشیم. شاخص‌های مورد ارزیابی در جدول ۳ آورده‌شده است.

جدول۳- تأثیر تغییرات وزن بر روی راه‌حل‌ها در الگوریتم بهینه‌سازی ازدحام ذرات

باکتری ارزیابی	W_1	W_2	W_3	تابع هدف
E.coli	۰/۵	۰/۳	۰/۲	۰/۷۸
	۰/۵	۰/۲۵	۰/۲۵	۰/۸۲
	۰/۵	۰/۲	۰/۳	۰/۹۳

در جدول ۴ پیچیدگی چند روش معروف انطباق چندگانه آورده شده است. همانطور که مشاهده می‌کنید در تمامی روش‌ها، پیچیدگی زمانی با توان دوم طول توالی‌ها (L) رابطه دارند. روش پیشنهادی به‌دلیل شکستن توالی‌ها به K زیر توالی، توالی‌ها را با پیچیدگی نه بیشتر از O(L^2) در راستای عمودی تقسیم می‌کند و بدون تحمیل هزینه اضافی، پیچیدگی زمانی حل مسأله را کاهش داده است. و چون در تطابق زیر رشته‌های حاصل، از روش‌های فراابتکاری ترکیبی استفاده می‌کند و با توجه به کوچک شدن ابعاد مسأله، فضای جستجوی راه‌حل‌های ممکن با سرعت بیشتر و دقت بالاتر به‌دست می‌آید. در این جدول L نشان‌دهنده طول توالی و N نشان‌دهنده تعداد توالی و k نشان دهنده تعداد زیر توالی‌های شکسته شده می‌باشد.

جدول ٤- پیچیدگی زمانی روش‌های تطابق چندگانه شناخته شده

نام روش	پیچیدگی زمانی
ClustalW	$O(N^2L^2)$
MUSCLE	$O(N^4 + NL^2)$
MAFFT	$O(N \log N) + O(NL^2)$
Dialign	$O(N^2 + L^3)$
Tlpso	$O(N^4 + NL^2)$
Halign-II	$O(NL^2)$
fame	$+O(\text{msa tools})O\left(KN\ \frac{L^2}{k^2}\right)$
روش پیشنهادی	$+NLO\left(N\ \frac{L^2}{k}\right)$

برای شکستن توالی‌ها، ابتدا روی تمامی توالی‌ها، هسته‌ها را می‌یابیم که پیچیدگی آن O(L) خواهد بود. پس از پیداکردن تشابهات، موقعیت آنها را استخراج کرده و آنها را با توجه به موقعیتی که روی یکی از توالی‌ها دارند مرتب می‌کنیم که این عمل دارای پیچیدگی O(L log L) است. در پایان با استفاده از برنامه‌نویسی پویا مسیرهای شکست عمودی را خواهیم یافت که پیچیدگی این بخش نیز O(L^2) خواهد بود. که در کل پیچیدگی تمام طول توالی O(L^2)می‌شود. که این میزان به‌دلیل شکستن توالی به k زیر توالی، به کسری از O(L^2) کاهش می‌یابد.

نتایج

برای ارزیابی صحت روش پیشنهادی، نتایج را با چندین روش پیشرو، تکرارشونده، فراابتکاری، تقسیم و غلبه مقایسه شده است. نتایج کیفیت دقت انطباق در جدول ۵ براساس معیار SP و نتایج افزایش سرعت انطباق در جدول ۶ خلاصه شده است.

با توجه به نتایج آزمایش‌ها در جدول ۶ الگوریتم پیشنهادی از نظر زمان و سرعت با سایر روش‌ها مقایسه می‌شود. لازم به ذکر است که اعداد موجود در جدول ۶ براساس دقیقه است. روش‌های FTLPSO و Dalign برای تکمیل فرایند انطباق (بیش از ۵ روز) به زمان بیشتری نیاز دارند. علاوه بر این، در بین سایر روش‌ها، روش Halign-II به علت مکانیزم موازی‌سازی و اجرای همروند چند پردازش با هم در یک محیط توزیع شده توانسته عملکرد خوبی را از خود نشان دهد. روش پیشنهادی نسبت به سایر روش‌ها سرعت را حدود ۳۰٪ افزایش داده است.

جدول ٥- مقایسه دقت انطباق روش پیشنهادی با سایر روش‌ها

دقت	روش پیشنهادی	Halign-II	FTLPSO	Dalign	muscle	Fame
Variola virus	۰/۹۳	۰/۹۰	۰/۸۱	۰/۸۱	۰/۹۳	۰/۹۲
Mycoplasma	۰/۷۲	۰/۶۹	۰/۶۵	۰/۶۸	۰/۷۰	۰/۷۱
Streptococcus	۰/۸۳	۰/۸۰	۰/۷۶	۰/۷۹	۰/۸۲	۰/۸۰
E.coli homolog	۰/۹۳	۰/۹۱	۰/۸۵	۰/۸۹	۰/۹۲	۰/۹۰
Sorangium	۰/۸۶	۰/۸۴	بالای دو هفته	۰/۸۲	۰/۸۶	۰/۸۴
titin	۰/۷۴	۰/۷۳	۰/۷۱	۰/۷۱	۰/۷۴	۰/۷۰
Menangit	۰/۷۳	۰/۷۱	۰/۶۹	۰/۷۰	۰/۷۴	۰/۷۱

جدول ٦- مقایسه سرعت انطباق روش پیشنهادی با سایر روش‌ها

سرعت	روش پیشنهادی	Halign-II	FTLPSO	Dalign	muscle	Fame
Variola virus	۰/۶	۰/۶	۲۰۰	۱۰۶	۹۰	۰/۹
Mycoplasma	۲/۷	۳	۸۰۰	۴۸۶	۱۷۵	۳
Streptococcus	۱۷/۳	۲۲	۱۵۶۰	۱۲۷۵	۱۰۱۷	۱۸
E.coli homolog	۲۱	۴۷	۵۱۵۰	۴۳۶۰	۴۵۶۰	۳۰
Sorangium	۷۹۰	۸۰۰	بالای دو هفته	۶۷۳۸	۷۸۸۰	۸۱۰
titin	۰/۸	۰/۹	۱۰	۷	۲	۱
Menangit	۵۶۰	۵۰۰	۲۹۰۰	۱۷۳۰	۱۳۵۰	۶۷۰

تشکر و قدردانی

از تمامی حمایت‌های دانشگاه آزاد اسلامی واحد شیراز و مسئولین محترم مجله علمی-پژوهشی "دانش و تندرستی در علوم پایه پزشکی" کمال تشکر و قدردانی را داریم.

References

1. Can T. Introduction to bioinformatics. InmiRNomics: MicroRNA Biology and Computational Analysis 2014:51-71. doi:10.1007/978-1-62703-748-8_4
2. Balech B, Monaco A, Perniola M, Santamaria M, Donvito G, Vicario S, Maggi G, Pesole G. DNA multiple sequence alignment guided by protein domains: The MSA-PAD 2.0 method. InViral Metagenomics 2018:173-80. doi:10.1007/978-1-4939-7683-6_13
3. DeBlasio D, Kececioglu J. Ensemble multiple alignment. InParameter Advising for Multiple Sequence Alignment 2017:85-102.
4. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 1994;22:4673-80. doi:10.1093/nar/22.22.4673
5. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic acids research 2004;32:1792-7. doi:10.1093/nar/gkh340
6. Katoh K, Misawa K, Kuma KI, Miyata T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research 2002;30:3059-66. doi:10.1093/nar/gkf436
7. Naznooshsadat E, Elham P, Ali SZ. FAME: fast and memory efficient multiple sequences alignment tool through compatible chain of roots. Bioinformatics 2020;36:3662-8. doi:10.1093/bioinformatics/btaa175
8. Zhang F, Qiao XZ, Liu ZY. A parallel smith-waterman algorithm based on divide and conquer. InFifth International Conference on Algorithms and Architectures for Parallel Processing, 2002. Proceedings 2002:162-9. doi:10.1109/ICAPP.2002.1173568
9. Wan S, Zou Q. HAlign-II: efficient ultra-large multiple sequence alignment and phylogenetic tree reconstruction with distributed and parallel computing. Algorithms for Molecular Biology 2017;12:25. doi:10.1186/s13015-017-0116-x
10. Morgenstern B. DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. Nucleic acids research 2004;32:W33-6. doi:10.1093/nar/gkh373
11. Rubio-Largo Á, Vega-Rodríguez MA, González-Álvarez DL. Hybrid multiobjective artificial bee colony for multiple sequence alignment. Applied Soft Computing 2016;41:157-68. doi:10.1016/j.asoc.2015.12.034
12. Du Y, He J, Du C. A novel binary particle swarm optimization for multiple sequence alignment. InInternational Conference on Intelligent Computing 2019:13-25. doi:10.1007/978-3-030-26969-2_2
13. Dabba A, Tari A, Zouache D. Multiobjective artificial fish swarm algorithm for multiple sequence alignment. INFOR: Information Systems and Operational Research 2020;58:38-59. doi:10.1080/03155986.2019.1629782
14. Moustafa N, Elhosseini M, Taha TH, Salem M. Fragmented protein sequence alignment using two-layer particle swarm optimization (FTLPSO). Journal of King Saud University-Science 2017;29:191-205. doi:10.1016/j.jksus.2016.04.007
15. Lalwani S, Kumar R, Gupta N. A novel two-level particle swarm optimization approach for efficient multiple sequence alignment. Memetic Computing 2015;7:119-33. doi:10.1007/s12293-015-0157-y
16. Gondro C, Kinghorn BP. A simple genetic algorithm for multiple sequence alignment. Genetics and Molecular Research 2007;6:964-82.
17. Angiuoli SV, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics 2011;27:334-42. doi:10.1093/bioinformatics/btq665
18. Moshiri N. ViralMSA: Massively scalable reference-guided multiple sequence alignment of viral genomes. Bioinformatics 2021;37:714-6. doi:10.1093/bioinformatics/btaa743
19. Morisse P, Marchet C, Limasset A, Lecroq T, Lefebvre A. Scalable long read self-correction and assembly polishing with multiple sequence alignment. Scientific reports 2021;11:1-3. doi:10.1038/s41598-020-80757-5
20. Lalwani S. Design and implementation of bi-level artificial bee colony algorithm to train hidden Markov models for performing multiple sequence alignment of proteins. International Journal of Swarm Intelligence 2021;6:48-64. doi:10.1504/IJSI.2021.114765
21. Pujari JJ, Pavan KK. Multiple sequence alignment based on enhanced brainstorm optimization algorithm with dynamic population size (EBSODP). Annals of the Romanian Society for Cell Biology 2021:10033-42.

Journal of Knowledge & Health
in Basic Medical Sciences 2021; Vol 16, No 1

DOI: 10.22100/jkh.v16i1.2512

Original Article

A New Multiple DNA and Protein Sequences Alignment Method based on Evolutionary Algorithms

Naznooshsadat Etminan (Ph.D. Student)[1], Elham Parvinnia (Ph.D.)[1*], Ali Sharifi-Zarchi (Ph.D.)[2]

1- Dept. of Computer Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran.
2- Dept. of Computer Engineering, Sharif University of Technology, Tehran, Iran.

Received: 19 February 2021, **Accepted:** 5 April 2021

Abstract:

Introduction: *The study of life and the detection of gene functions is an important issue in biological science. Multiple sequences alignment methods measure the similarity of DNA sequences. Nonetheless, when the size of genome sequences is increased, we encounter with the lack of memory and increasing the run time. Therefore, a fast method with a suitable accuracy for genome alignment has a significant impact on the analysis of long sequences.*
Methods: *We introduce a new method in which, it first divides each sequence into short sequences. Then, it uses evolutionary algorithms to align the sequences.*
Results: *The proposed method has been evaluated in seven datasets with different number of nucleotides per DNA sequence (18,000 to 14 million) and compared to five popular multiple sequences alignment methods. The highest accuracy for the variola bacterium dataset is 93% and the highest alignment rate is 0.6 per minute for this bacterium.*
Conclusion: *Most multiple alignment methods in short sequences or datasets with only a few sequences have good accuracy while require high computational time for longer sequences. The proposed algorithm overcomes this drawback by aligning long sequences in an acceptable time and maintaining accuracy as well as optimal memory usage.*

Keywords: Multiple Sequence alignment, Complete Genome Data, Sequence Division, Evolutionary Algorithms.

Conflict of Interest: No
***Corresponding author:** E Parvinnia, **Email**: eparvinnia@yahoo.com*

Citation: Etminan N, Parvinnia E, Sharifi-Zarchi A. A new multiple dna and protein sequences alignment method based on evolutionary algorithms. Journal of Knowledge & Health in Basic Medical Sciences 2021;16(1):13-20.