وزارت علوم، تحقیقات و فناوری
دانشگاه تحصیلات تکمیلی علوم پایه
گاوازنگ - زنجان

انطباق چند توالی بیولوژیکی با استفاده از الگوریتم جستجوی ممنوعه

پایان‌نامهٔ کارشناسی‌ارشد

محمدحسین نظرزاده

استاد راهنما: **دکتر مهدی وثیقی**

فروردین ۱۳۹۳


بسم الله الرحمن الرحیم

تقدیم به پروفسور یوسف ثبوتی

که حضور گرانقدر و اندیشه پربارشان، مایه امید و سرمشق شاگردانی است که در مکتب ایشان پرورش یافته‌اند.

تشکر و قدردانی

در اینجا از همه‌ی دوستانم که در طی دوران تحصیل لحظات فراموش نشدنی و لذت‌بخشی را با هم سپری نمودیم تشکر می‌کنم.

چکیده

انطباق چند توالی به‌طور گسترده‌ای در زمینه‌ی بیوانفورماتیک مورد استفاده واقع شده است که از پیش‌بینی ساختار، تجزیه و تحلیل بیولوژیکی و شبیه‌سازی فیلوژنتیکی می‌توان به‌عنوان برخی از کاربردهای آن نام برد. طیف وسیعی از روش‌ها به‌منظور حل مسئله‌ی انطباق چند توالی ارائه گردیده است با این حال برخی از آن‌ها به پیش‌نیازهایی جهت یافتن انطباق بهینه احتیاج دارند و یا ممکن است از معایبی همچون پیچیدگی محاسباتی بالا و زیاد بودن حافظه‌ی مورد نیاز رنج ببرند. این عوامل باعث شده است که این روش‌ها تنها در مورد تعداد محدودی از توالی‌ها اعمال گردند.

در این پایان‌نامه ما از روش جستجوی ممنوعه با هدف حل مسئله‌ی انطباق چند توالی سود برده‌ایم. جستجوی ممنوعه یک روش بهینه‌سازی است که با استفاده از ویژگی‌های حافظه‌ی انطباقی به جستجو و یافتن راه‌حل مناسب می‌پردازد. ویژگی‌های حافظه‌ی انطباقی می‌تواند الگوریتم را از فضای بهینه‌ی محلی دور و فضای جواب را به شکل کارآمدی جستجو نماید.

به‌منظور دست‌یابی به نتایج بهتر، روش‌های مختلف تولید همسایگی، عرضه‌ی لیست ممنوعه‌ای با ساختار جدید و استفاده از راهکار نوینی جهت انجام راهکار تنوع‌بخشی ارائه شده است. عملکرد روش پیشنهادی با استفاده از پایگاه داده‌ی *BALiBASE* مورد بررسی قرار گرفته است. افزون براین نتایج به‌دست آمده از ترکیب الگوریتم پیشنهادی با روش *ClustalW* (روشی تدریجی) نشان دهنده‌ی عملکرد بسیار مطلوبی در زمینه‌ی حل مسئله‌ی انطباق چند توالی بوده است.

واژه‌های کلیدی: *توالی‌های بیولوژیکی، انطباق چند توالی، جستجوی ممنوعه*

فهرست

فهرست تصاویر

فهرست جداول

پیش‌گفتار

انطباق چند توالی به‌عنوان راهکاری مفید در زمینه‌ی تجزیه و تحلیل توالی‌های بیولوژیکی مطرح شده است. با اثبات NP -کامل بودن مسئله‌ی انطباق چند توالی دست‌یابی به انطباقی بهینه از مهم‌ترین دغدغه‌ها در حیطه‌ی بیوانفورماتیک بوده به‌طوری‌که تاکنون مطالعات فراوانی در این راستا صورت پذیرفته است.

اگرچه روش‌های سریعی جهت حل مسئله‌ی انطباق چند توالی ارائه گردیده است اما کارایی این روش‌ها در یافتن انطباقی بهینه با افزایش طول و تعداد توالی‌های شرکت کننده در انطباق کاهش یافته است. از سویی دیگر استفاده از روش‌های تکراری، هرچند زمان نسبتا زیادی را صرف به‌دست آوردن انطباق بهینه کرده‌اند، تا حد قابل قبولی توانسته است معایب سایر روش‌ها را برطرف نماید.

با توجه به مطالب بیان شده استفاده از الگوریتم جستجوی ممنوعه می‌تواند در زمینه‌ی حل مسئله‌ی انطباق چند توالی راهگشا باشد. افزون براین ساختار الگوریتم جستجوی ممنوعه به‌گونه‌ای است که به‌کارگیری آن جهت انطباق توالی‌ها با شروع از جواب اولیه‌ی انطباق یافته مؤثر واقع شده است.

فصل اول

مقدمه

انطباق (تطابق) چند توالی[۱] (*MSA*) از دیرباز به‌عنوان مسئله‌ای مهم در زمینه‌ی طراحی الگوریتم مطرح بوده است. کشف ژنوم انسان و نیاز به‌وجود آمده در زمینه‌ی تطابق توالی‌های ژنتیکی بیش‌ازپیش بر اهمیت موضوع افزوده است تا آنجا که اگر چه بیش از ۳۰ سال از عمر اولین الگوریتم‌ها در این زمینه نمی‌گذرد اما همچنان به‌طور گسترده‌ای در زمینه‌ی زیست‌شناسی مولکولی مورد استفاده قرار گرفته‌اند. انطباق چند توالی دارای گستره‌ی وسیعی از کاربردها است که مهم‌ترین آن‌ها عبارتند از: پیدا کردن ویژگی‌ها و خصیصه‌های اصلی در میان توالی‌های بیولوژیکی، یافتن چرخه و سیر تکامل ژنتیکی در طول زمان بر اساس انطباق توالی‌های هم‌خانواده، شناسایی توالی‌های متعلق به یک خانواده، پیش‌بینی ساختارهای دوم و سوم توالی‌ها، تجزیه و تحلیل ساختارهای فیلوژنتیکی[۲] و جستجو در پایگاه داده‌های عظیم توالی.

بسیاری از تلاش‌ها و مطالعات صورت گرفته در زمینه‌ی انطباق چند توالی بر مبنای این اصل استوار

---

۱ Multiple Sequence Alignment

۲ Phylogenetic

است که این مسئله به‌عنوان یک مسئله‌ی NP - کامل[۱] [۱] شناخته شده است.

جهت به‌دست آوردن بهترین انطباق ممکن بر روی چند توالی الگوریتم‌هایی ارائه گردیده است اما پیچیدگی محاسباتی آن‌ها با افزایش طول و تعداد توالی‌های شرکت کننده در انطباق به‌صورت نمایی افزایش یافته است که همین امر استفاده از آن‌ها را تنها در مورد تعداد محدودی از توالی‌ها مقرون به‌صرفه کرده است. به‌منظور رفع این مشکل روش‌های دیگری جهت حل مسئله‌ی انطباق چند توالی به‌کار گرفته شده است که یکی از پرکاربردترین آن‌ها روش تطابق تدریجی[۲] است. این روش ابتدا به انطباق دو دنباله‌ای می‌پردازد که بیش‌ترین شباهت ساختاری را دارا هستند. پس از آن انطباق با اضافه کردن توالی‌های باقی‌مانده به‌صورت پی‌درپی ادامه یافته است. روش تطابق تدریجی بر مبنای الگوریتم حریصانه[۳] ارائه شده است و خطای انطباق یک بار در طول روند اجرا محاسبه می‌شود و قابلیت اصلاح شدن ندارد. این روش دارای کارایی بالایی از لحاظ سرعت اجرا است هرچند در نهایت به یک انطباق محلی منجر شده است.

الگوریتم‌های فراابتکاری[۴] دسته‌ی دیگری از روش‌ها هستند که جهت انطباق چند توالی مورد استفاده قرار گرفته‌اند. در این الگوریتم‌ها جواب اولیه بر مبنای یک تابع ارزش‌گذاری مشخص به‌صورت تکراری بهبود یافته است. الگوریتم شبیه‌سازی حرارتی[۵] و الگوریتم ژنتیک[۶] از جمله‌ی این روش‌هاست که تا اکنون جهت انطباق چند توالی به‌کار رفته‌اند. با توجه به نبود الگوریتمی قطعی جهت حل مسئله‌ی انطباق چند توالی در زمان چندجمله‌ای، استفاده از الگوریتم‌های فراابتکاری می‌تواند راهگشا باشد (الگوریتم‌هایی که دست‌یابی به بهترین جواب ممکن را تضمین نمی‌کنند اما تا حدود بسیار زیادی به

---

۱ NP-Complete

۲ Progressive Method

۳ Greedy Algorithm

۴ Metaheuristic

۵ Simulated annealing algorithm

۶ Genetic algorithm

جواب بهینه نزدیک می‌شوند).

جستجوی ممنوعه[1] الگوریتمی اکتشافی است که بر اساس تکرار در هر مرحله به جواب بهینه نزدیک‌تر می‌شود. در این الگوریتم ابتدا جواب اولیه انتخاب و سپس جواب‌های همسایه از روی آن ساخته می‌شود (بر مبنای یک روش مشخص). آن‌گاه بهترین همسایه بر اساس تابع ارزش‌گذاری تعیین و جایگزین جواب فعلی می‌گردد. الگوریتم جستجوی ممنوعه به‌طور گسترده‌ای جهت حل مسائل موجود در زمینه‌های مختلف علمی به‌کار رفته است. این الگوریتم همچنین جهت انطباق چند توالی مورد استفاده قرار گرفته است. الگوریتم جستجوی ممنوعه به راحتی به‌صورت موازی اجرا و همین امر کاهش زمان اجرای الگوریتم و محبوبیت آن را به‌دنبال داشته است.

افزون بر مباحث مطرح شده در زمینه انطباق چند توالی، موضوعی که بیش‌ازبیش بر اهمیت حل این مسئله افزوده است یافتن الگوریتمی است که علاوه بر ارائه‌ی یک انطباق خوب، حاوی اطلاعات بیولوژیکی مربوط به توالی‌ها نیز باشد. به‌عبارت دیگر استفاده از الگوریتم‌هایی که بدون در نظر گرفتن اطلاعات بیولوژیکی، صرفا ارائه کننده‌ی یک انطباق باشند نمی‌تواند مفید واقع گردد. امروزه از داده‌ها و اطلاعات پیچیده‌ی مربوط به ساختار و عملکرد پروتئین‌ها جهت به‌دست آوردن یک انطباق مفید استفاده شده است. به‌کارگیری این اطلاعات جدید و اضافی نتایج واقع بینانه‌تری را به‌دنبال داشته است.

با توجه به مطالب بیان شده این پایان‌نامه به‌دنبال ارائه‌ی روشی مفید جهت حل مسئله‌ی انطباق چند توالی است. جهت نائل شدن به این هدف از الگوریتم جستجوی ممنوعه استفاده شده است. همچنین توابع ارزش‌گذاری متفاوت و جدید مورد بررسی قرار گرفته‌اند که بر اساس آن می‌توان اطلاعات مفیدتری را در انطباق دخیل نمود.

فصل دوم این پایان‌نامه به توضیح مفاهیم بنیادی در زمینه‌ی انطباق چند توالی پرداخته است. همچنین

---

[1] Tabu search

به اختصار در رابطه با انطباق چند توالی بحث و به‌صورت اجمالی الگوریتم‌های موجود در این زمینه مطرح شده است.

فصل سوم به معرفی الگوریتم جستجوی ممنوعه و پارامترهای وابسته به آن اختصاص یافته است. مروری بر کارهای گذشته در زمینه‌ی انطباق چند توالی با استفاده از الگوریتم جستجوی ممنوعه از سایر مباحث مطرح شده در این فصل است.

در فصل چهارم روش پیشنهادی مطرح و نتایج به‌دست آمده از آن با نتایج حاصل از سایر روش‌های موجود مقایسه شده است. همچنین در این فصل درباره مزیت استفاده از روش پیشنهادی به بحث پرداخته شده است.

فصل پنجم نیز به نتیجه‌گیری و بیان کارهای آینده در زمینه‌ی به‌کارگیری الگوریتم جستجوی ممنوعه جهت انطباق چند توالی بیولوژیکی اختصاص یافته است.

فصل دوم

مروری بر مسئله‌ی انطباق چند توالی

در این فصل مفاهیم بیولوژیکی مرتبط با مسئله‌ی انطباق چند توالی بیان شده است. ارائه‌ی توابع ارزش‌گذاری متفاوت از سایر مباحث مطرح شده در این فصل است. افزون براین مروری بر الگوریتم‌های موجود در زمینه‌ی حل مسئله‌ی انطباق چندین توالی صورت پذیرفته است.

موجودات زنده‌ی مختلف از لحاظ بیولوژیکی دارای شباهت‌های قابل ملاحظه‌ای می‌باشند. به‌عبارت دیگر دارای ساختارهای ژنتیکی مشابهی هستند. بررسی‌ها بر روی داده‌های به‌دست آمده از توالی‌های بیولوژیکی نشان می‌دهد که اشتراک ژنی بین دو پستاندار می‌تواند به میزان ۹۹٪ افزایش یابد. انسان‌ها و میوه‌ها هرچند در دو دسته‌ی مختلف جای دارند اما دارای حداقل ۵۰٪ شباهت ژنتیکی می‌باشند. این حقایق قابل توجه تا حد زیادی از طریق تجزیه و تحلیل توالی‌های زیستی کشف شده‌اند. انطباق چند توالی بیولوژیکی یکی از ابتدایی‌ترین و اساسی‌ترین کارهای صورت گرفته در زمینه بیوانفورماتیک و آنالیز توالی‌های ژنتیکی است.

در اوایل دهه‌ی ۱۹۷۰ توالی اسید دئوکسی‌ریبونوکلئیک *(DNA)*[۱] با استفاده از روش‌های دشوار، بر

[۱] Deoxyribonucleic Acid

اساس کروماتوگرافی دوبعدی به‌دست آمد. با این وجود هنوز تعداد اندکی از توالی‌ها شناخته شده بود و هر توالی به‌صورت مجزا از سایر توالی‌ها مورد بررسی قرار می‌گرفت. در اواخر دهه‌ی ۷۰ والتر گیلبرت[۱] و فردریک سانگر[۲] توانستند بر اساس روش‌های تخریب شیمیایی و سنتز آنزیمی توالی *DNA* را به‌دست آورند. این کار جایزه نوبل در سال ۱۹۸۰ را برای آن‌ها به ارمغان آورد. از آن پس توالی‌ها بر اساس روش‌های جدیدتری از جمله روش‌های مبتنی بر رنگ، میکرو آرایه‌ها و اشعه‌ی ایکس به‌دست آمدند. امروزه حدود ۲۰ میلیون توالی در جهان شناخته شده است که این تعداد همه‌روزه در حال افزایش است. این توالی‌ها توسط پایگاه‌های مختلف نگهداری و مورد استفاده قرار می‌گیرد.

با توجه به حجم عظیم و متنوعی از توالی‌ها، تجزیه و تحلیل هر توالی به‌صورت دستی و مجزا از بقیه غیرممکن می‌نماید. بنابراین در وهله‌ی اول نیاز به دسته‌بندی توالی‌ها بیش از سایر نیازها به چشم می‌آید. آن‌گاه آنالیز بر روی توالی‌های جای گرفته در هر گروه صورت می‌پذیرد و مشخصه‌ها و ویژگی‌های مشترک استخراج می‌گردد. در این راستا اولین گام یافتن روشی بهینه جهت تشخیص و گروه‌بندی توالی‌ها بر اساس ویژگی‌های ساختاری آن‌ها می‌باشد. سپس با استفاده از الگوریتم‌ها و تکنیک‌های موجود در زمینه‌ی انطباق، بهترین تطابق موجود بین توالی‌ها به‌دست می‌آید که بر اساس آن می‌توان به تجزیه و تحلیل توالی‌های بیولوژیکی پرداخت.

به‌طور کلی در زمینه بررسی توالی‌ها دو روش بیش از بقیه عمومیت یافته است که این مسئله از برخورد و نگرش متفاوت دانشمندان بیوانفورماتیک نشأت گرفته است. اولین روش بر استفاده از سازوکارهای دقیق الگوریتمی جهت به‌دست آوردن بهترین انطباق موجود بین توالی‌ها تمرکز می‌کند. در نتیجه، اجرای این روش منجر به آشکار شدن روابط و تشابهات بیولوژیکی بین توالی‌ها شده که از آن به‌منظور دسته‌بندی و سازماندهی پایگاه داده توالی‌ها استفاده گردیده است. افزون براین پیش‌بینی ساختارهای دوم و سوم توالی‌های بیولوژیکی از مزایای به‌کارگیری این روش است.

---

[۱] Walter Gilbert

[۲] Frederick Sanger

روش دوم درپی شناسایی روند تکامل بیولوژیکی توالی‌هاست. این توالی‌ها به یک خانواده تعلق داشته و ممکن است همه و یا برخی از ویژگی‌های آن‌ها شناخته شده باشد. هدف از به‌کارگیری این روش، کشف ساختارهای بیولوژیکی توالی‌هاست که باعث به‌وجود آمدن ویژگی‌ها و خصوصیات مشترک بین توالی‌ها شده است. در نتیجه می‌توان از نتایج به‌دست آمده جهت تجزیه و تحلیل تکامل، آنالیز عملکرد و ساختار توالی‌ها و کشف و طراحی داروها استفاده نمود. به‌عنوان مثال، طراحان دارو بر اساس انطباق به‌دست آمده از توالی یک ویروس جدید و توالی‌های موجود در پایگاه داده‌های ژن، می‌توانند ساختار ویروس را کشف کرده و به ساخت دارو مبادرت ورزند.

درنهایت، دو دیدگاه مطرح شده باعث به‌وجود آمدن روش‌های مختلفی جهت انطباق چندین توالی شده است. توسعه‌ی الگوریتم‌های انطباق به‌نوبه‌ی خود به دانشمندان اجازه داده است با سرعت و دقت بیشتر و در زمان کمتر به تجزیه و تحلیل توالی‌ها بپردازند.

۱.۲ مفاهیم بنیادی زیستی

مشخصه‌ی اصلی یک موجود زنده *DNA* آن می‌باشد که حاوی ژنوم آن موجود است. به‌عبارت دیگر *DNA* مجموعه‌ای از اطلاعات ژنتیکی را در درون خود ذخیره کرده است. *DNA* از هزاران ژن تشکیل شده است. هر ژن حاوی اطلاعاتی از چگونگی ساخت یک واحد پروتئینی است که به‌عنوان اجزای سازنده‌ی یک سلول عمل می‌کنند. افزون براین انجام برخی از سازوکارهای درون سلولی بر عهده‌ی واحدهای پروتئینی است. از آنجا که *DNA* دربردارنده‌ی اطلاعات ژنتیکی سلول است قبل از تقسیم شدن سلول باید تکثیر شود. از این فرآیند با عنوان همانندسازی[۱] یاد شده است. هنگام ساخت پروتئین،

[۱] Duplication

ژن مربوط به *RNA*[۱] رونویسی[۲] می‌شود. سپس بخش‌های غیرکدگذاری شده‌ی *RNA* *(Intron)* حذف و *RNA* به خارج از هسته‌ی مولکول منتقل می‌گردد. پروتئین‌ها در خارج از هسته بر اساس کدون‌های (کدهای سه تایی) موجود بر روی *RNA* ساخته می‌شوند. بنابراین توالی *DNA* نقشی تعیین کننده در ساختار و عملکرد پروتئین بر عهده دارد.

DNA ۱.۱.۲

اسید دئوکسی‌ریبونوکلئیک *(DNA)* ذخیره کننده و حافظ اطلاعات ژنتیکی سلول است. *DNA* پلیمری با اندازه‌ای بزرگ بوده که از نوکلئوتیدها[۳] ساخته شده است. اسیدهای نوکلئیک[۴] نامی است که به این زنجیره‌ی بلند از مولکول‌ها اتلاق می‌شود. سه جزء عمده‌ی موجود در نوکلئوتیدها عبارتند از: اسید فسفریک، مونوساکاریدهای پنج‌کربنی (ریبوز[۵] و دئوکسی‌ریبوز[۶]) و یك ترکیب حلقوی نیتروژن‌دار بازی (پورین[۷] یا پیریمیدین[۸]).

پورین‌های موجود در نوکلئوتیدها به‌طور عمده از آدنین *A*[۹] و گوانین *G*[۱۰] تشکیل شده است. افزون

---

۱ Ribonucleic Acid

۲ Transcription

۳ Nucleotide

۴ Nucleic Acid

۵ Ribose

۶ Deoxyribose

۷ Purine

۸ Pyrimidine

۹ Adenine

۱۰ Guanine