دانشگاه شهید مدنی آذربایجان
وزارت علوم، تحقیقات و فناوری
دانشگاه شهید مدنی آذربایجان
دانشکده علوم پایه
گروه ریاضی کاربردی

پایان نامه برای دریافت درجه کارشناسی‌ارشد
رشته ریاضی کاربردی

عنوان

کاربرد بهینه سازی ترکیبیاتی در زیست شناسی محاسباتی

استاد راهنما
دکتر علیرضا غفاری حدیقه

استاد مشاور
دکتر نادر چاپارزاده

پژوهشگر
رضا مددی حور

بهمن /۱۳۹۳
تبریز / ایران

تقدیم به:

پدر و مادر عزیزم و خواهرزاده‌ام هلنا عشق پاک زندگی‌ام

سپاس گزاری...

و مَن لَمْ یَشْکُر الْمَخلوق لَمْ یَشْکُر الْخالِقْ.

سپاس و ستایش پروردگار متعال را که به اینجانب توفیق تلاش در راه کسب علم و دانش را عطا فرمود. امیدوارم بتوانم آموخته‌هایم را در راه پیشرفت علمی وطن خویش مورد استفاده قرار دهم.

در آغاز وظیفه‌ی خود می‌دانم از زحمات بی دریغ استاد راهنمای خود، جناب آقای دکتر **علیرضا غفاری حدیقه**، صمیمانه تشکر و قدردانی نمایم که قطعاً بدون راهنمایی‌های ارزنده‌ی ایشان این مجموعه به انجام نمی‌رسید.

از جناب آقای دکتر **نادر چاپارزاده** که زحمت مطالعه و مشاوره این رساله را تقبل فرمودند و در آماده سازی این رساله، به نحو احسن اینجانب را مورد راهنمایی قرار دادند، کمال امتنان را دارم.

همچنین با سپاس بی‌دریغ خدمت سرکار خانم **زرندی** و جناب آقای **حسین مصطفایی** که مرا صمیمانه و مشفقانه یاری داده‌اند.

و از جناب آقای دکتر **بهروز خیرفام** سپاسگزارم که قبول زحمت فرموده و داوری این تحقیق را برعهده گرفتند.

با تقدیر و درود فراوان خدمت پدر و مادر بسیار عزیز، دلسوز و فداکارم که پیوسته جرعه نوش جام تعیلم و تربیت، فضیلت و انسانیت آنها بوده ام و همواره چراغ وجودشان روشنگر راه من در سختی‌ها و مشکلات بوده است و تشکر می‌کنم از برادر و خواهر عزیزم، به پاس کمک‌های بی‌دریغ و دلگرمی‌های بی‌پایان‌شان، که بهترین پشتیبان من بودند.

رضامددی حور

بهمن ۱۳۹۳

تبریز - ایران

فهرست مطالب

چکیده

درک چگونگی کدگذاری از اطلاعات مشخص‌کننده زمان و مکان رونویسی یک ژن به محصول پروتئینی، هدف اصلی در زیست‌شناسی مولکولی است. پروتئین‌های واسطه معروف به فاکتورهای رونویسی این فرآیند را به وسیله‌ی تعامل با DNA های سلول و ماشین‌های رونویسی تسهیل می‌بخشند. این از اهمیت بسزایی در شناسایی همه توالی‌های محل‌های اتصال فاکتور رونویسی درDNA مشخص برخوردار است. در این پایان‌نامه مسأله محاسباتی مربوط به کشف موتیف بررسی و مطالعه می‌شود. در این زمینه یک مجموعه از توالی‌های معلوم که شامل محل‌های اتصال از یک فاکتور رونویسی خاص هستند داده شده است، هدف این است تا موقعیت‌های آنها را شناسایی کنیم. در اینجا یک چارچوب بهینه‌سازی ترکیبیاتی برای موتیف یابی که از هردو روش شاخه‌بری گراف و فرمول‌بندی برنامه‌ریزی خطی عددصحیح استفاده می‌کند بررسی شده است. همچنین روشی برای شناسایی موتیف‌هایی با معنی‌داری آماری معرفی می‌شود. الگوریتم موردنظر برای شماری از مجموعه داده‌های زیستی و داده‌های مصنوعی اعمال شده و به خوبی اجرا می‌شود. همچنین قابل اجرا بودن این الگوریتم برای انواع دیگری از مسأله تشخیص محل اتصال DNA مانند ردپانگاری فیلوژنتیک، فرمول‌بندی موتیف‌های ظریف و مسأله موتیف‌های چندگانه بررسی و مطالعه می‌شود. نتایج نشان می‌دهند که ترکیب نظریه گراف و رویکرد برنامه‌ریزی ریاضی، می‌تواند پایه مناسبی برای روش‌های مؤثر و قدرتمند در پیداکردن کاربردهای مختلف موتیف یابی باشد

کلمات کلیدی: موتیف یابی، برنامه‌ریزی عددصحیح، بهینه سازی ترکیبیاتی، نظریه گراف

پیشگفتار

رشته زیست شناسی محاسباتی در ۲۰ سال گذشته رشد فوق العاده ای را تجربه کرده است. کشف تقریبی الگوهای تکرارشونده یا موتیف‌ها در توالی‌های زیستی یک مساله مهم و مورد مطالعه در زیست شناسی مولکولی محاسباتی است. اغلب کاربردهای موتیف یابی، هنگام شناسایی سیگنال‌های تنظیمی مشترک در داخل توالی‌های DNA یا عناصر ساختاری و عملکردی مشترک در داخل توالی‌های پروتئین آشکار می‌شود. موتیف‌ها معمولاً با آنالیز آزمایشگاهی تعیین می‌شوند که فرآیندی بسیار وقت‌گیر و دشوار است. پیش‌گویی کامپیوتری موتیف‌ها بسیار امید بخش است چرا که پتانسیل جایگزینی با حجم عظیمی از آنالیزهای آزمایشگاهی را دارد.

در پیداکردن موتیف‌ها می‌توان به دو زیرمسأله اشاره کرد که عبارتند از کشف و نمایش موتیف‌ها[۸۲]. مسأله کشف موتیف عبارت است از پیدا کردن نمونه‌های موتیف در یک مجموعه از توالی‌های شناخته شده(اغلب به صورت آزمایشگاهی) که شامل محل‌هایی برای یک فاکتور مشترک یا موتیف‌ها هستند.

اولین مطالعات علمی روی مسأله موتیف یابی در سال ۱۹۹۰ توسط لاورنس و ریلی [۱] [۵۵] شروع شد که یک الگوریتم بیشینه‌سازی مورد انتظار را برای شناسایی و توصیف محل‌های مشترک در توالی بیوپلیمرهای همردیف نشده مورد بررسی قرار گرفت. لاورنس [۲] و همکاران در سال ۱۹۹۳ برای تشخیص سیگنال‌های ظریف توالی روش نمونه گیری موتیف گیبس در همردیفی چندگانه توالی را ایجاد کردند. بایلی و الکان [۳] [۱۱]در سال ۱۹۹۵ دانشی از موتیف‌های چندگانه در بیوپلیمرها را با استفاده از بیشینه‌سازی انتظار مورد بررسی قرار دادند. در سال

[۱] Lawrence and Reilly
[۲] Lawrence
[۳] Bailey and Elkan

۱۹۹۸ توسط برازما [۱] و همکاران [۱۳] روش‌هایی برای کشف خودکار الگوها در توالی‌های زیستی پیشنهاد شد. درهمان سال ریگوتسوس و فلوراتوس [۲] [۷۵]کشف الگوی ترکیبی در توالی‌های زیستی را بررسی کردند. هرتز و استورمو [۳] [۳۹]در سال ۱۹۹۹ شناسایی الگوهای پروتئین و DNA با معنی‌داری آماری را با استفاده از همردیفی چندگانه توالی مورد بررسی قرار دادند. در سال ۱۹۹۹ تومپا [۴] [۸۸]یک روش دقیق برای یافتن موتیف‌های کوتاه در توالی‌ها ارائه کرد و آن را برای مسأله محل اتصال ریبوزوم به‌کار برد. ورکمن و استورمو [۵] [۹۴]در سال ۲۰۰۰ یک روش برای کشف محل‌های اتصال فاکتور رونویسی با ویژگی های بهبود یافته ارائه کردند. پوزنر و اسزی [۶] [۷۱]اولین کسانی بودند که در سال ۲۰۰۰ روش‌های ترکیبیاتی را برای موتیف‌یابی اتخاذ کردند. آنها روش‌های ترکیبیاتی را برای پیدا کردن سیگنال‌های ظریف در توالی‌های زیستی به کار بردند. در سال ۲۰۰۱ لیو [۷] و همکاران [۵۴] کشف موتیف‌های محافظت‌شده DNA در نواحی تنظیمی بالادستی از ژن‌های هم‌بیان را مورد بررسی و مطالعه قرار دادند. اسکین و پوزنر[۸] [۲۶]در سال ۲۰۰۲ یافتن الگوهای تنظیمی مرکب در توالی‌های DNA را مورد بررسی قرار دادند. در سال ۲۰۰۳ سینها و تومپا [۹] [۷۹]برنامه‌ای با نام YMF را برای کشف موتیف طراحی کردند که برای تشخیص محل اتصال فاکتور رونویسی از روش‌های آماری استفاده می‌کرد. فریس [۱۰] و همکاران[۲۷] در سال ۲۰۰۴ پیدا کردن عناصر کارکردی توالی به وسیله همردیفی چندگانه موضعی را مورد بررسی و مطالعه قرار دادند. تومپا و همکاران [۸۶] در سال ۲۰۰۵ نشان دادند مسائل زیستی که به وسیله موتیف‌یابی بررسی می‌شوند، پیچیده و متنوع‌اند و هیچ روش منحصر به فردی نیست که بتواند همه آنها را به طور کامل حل کند.

دو گروه از الگوریتم‌ها برای موتیف‌یابی که در کارهای فوق‌الذکر به صورت گسترده استفاده شده اند عبارتند از:

[۱] Brazma
[۲] Rigoutsos and Floratos
[۳] Hertz and Stormo
[۴] Tompa
[۵] Workman and Stormo
[۶] Pevzner and Sze
[۷] Liu
[۸] Eskin and Pevzner
[۹] Sinh and Tompa
[۱۰] Frith

روش‌های جستجوی احتمالاتی که مبتنی بر نمایش ماتریس‌های امتیازدهی اختصاصی موقعیت (PSSM) هستند و روش‌های ترکیبیاتی که مبتنی بر نمایش انواعی از توالی‌های توافقی می‌باشند. هر دو گروه از الگوریتم‌ها با یک تابع امتیازدهی خاص ظاهر می‌شوند[۸۲، ۶۸]. اکثر مسائل موتیف‌یابی از جمله بهینه سازی امتیاز متوسط محتوای اطلاعات یا امتیاز مجموع جفت‌ها(SP) «NP-سخت» است. در سال ۲۰۰۴ اوسادا [۱] و همکاران با مقایسه روش‌های امتیازدهی نشان دادند که امتیازدهی SP یکی از معقول‌ترین طرح‌ها برای ارزیابی حفاظت موتیف‌هاست [۶۷] . که در مقالات کاریلو [۲] و لیپمن [۳] و اسکولر [۴] وهمکاران این امر مشهود است [۱۶، ۷۸]. در این پایان‌نامه یک چارچوب بهینه سازی ترکیبیاتی چند بعدی برای مسأله موتیف‌یابی مدل شده به وسیله همردیفی چندگانه موضعی توالی با طرح امتیازدهی SP ، بررسی و مطالعه می‌شود که در آن هدف پیدا کردن یک بیشینه خوشه از لحاظ وزن در یک گراف Nقسمتی است. از انواع مختلف موجود مسأله موتیف‌یابی در این پایان‌نامه اکثر این مسائل بررسی و مطالعه خواهند شد[۴۹].(۱)نمونه ساده که در آن هر توالی در مجموعه داده مفروض شامل دقیقاً یک نمونه موتیف است.(۲) نمونه متجاوز که در آن امکان دارد بیش از یک نمونه در برخی از توالی‌ها موجود باشد.(۳)نمونه معیوب که در آن امکان دارد یک نمونه موتیف در هر توالی رخ ندهد.(۴)نمونه‌های چندتایی که در آن امکان دارد توالی‌ها شامل بیش از یک نمونه موتیف مشترک باشند.

این پایان‌نامه درشش فصل تنظیم شده است. در فصل اول برخی مفاهیم و تعاریف پایه‌ای از زیست شناسی و ریاضی، که ممکن است در ابتدا برای خواننده ناآشنا به نظر برسد آمده است. در فصل دوم به معرفی مسائل بهینه سازی ترکیبیاتی پرداخته و پس از معرفی برنامه ریزی خطی، فنون حل مسائل بهینه سازی ترکیبیاتی را بیان کرده و به تحلیل پیچیدگی حل این مسائل می‌پردازیم. در فصل سوم به بررسی موتیف یابی در رشته های DNA یا پروتئین خواهیم پرداخت. در فصل چهارم رویکرد بهینه سازی ترکیبیاتی را برای موتیف یابی بیان و تحلیل می‌کنیم.

[۱] Osada
[۲] Carillo
[۳] Lipman
[۴] Schuler

در فصل پنجم با انتخاب ۱۶ نمونه از داده‌های باکتری ایشریشاکولای و داده‌های مصنوعی نتایجی را با استفاده از نرم‌افزارهای زیستی و ریاضی بدست می‌آوریم. در فصل ششم نتیجه گیری کلی از پایان‌نامه و پیشنهاداتی برای کارهای آینده آورده شده است.

فصل ۱
مفاهیم و تعاریف اولیه

زیست شناسی مولکولی محاسباتی یا به طور ساده‌تر زیست شناسی محاسباتی یک زمینه تحقیقاتی است که جواب مسایل محاسباتی ناشی از زیست شناسی مولکولی را مطالعه می‌کند. سؤالات در این رشته اغلب در انواعی از نظریه گراف، جریان‌های شبکه، ترکیبیات، مسائل برنامه‌ریزی خطی و عدد صحیح مطرح می‌شود. همچنین قابلیت بی پایان زیست شناسی محاسباتی در ارائه مسائل ترکیبیاتی جالب، برای جامعه تحقیق درعملیات و برنامه‌ریزی ریاضی، بسیار جذاب شده است. در واقع بسیاری از مسائل زیست شناسی محاسباتی می‌توانند به عنوان مسائل بهینه سازی مطرح شده و توسط فنون استاندارد بهینه سازی حل شوند.

روش معمول برای مقابله با یک مسأله زیست شناسی محاسباتی به شرح زیر است: ابتدا یک تحلیل مدل‌بندی انجام می‌شود به این ترتیب که فرآیندهای زیستی تحت بررسی، در داخل یک یا چند مورد ترکیبیاتی(همچون گراف‌ها، مجموعه‌ها، آرایه‌ها و ...)ترسیم می‌شوند. بعد از مرحله مدل‌بندی، سؤال اصلی درباره داده‌های زیستی به یک سؤال ریاضی در مواردی از نمایش‌های انتخاب شده تبدیل می‌شود[۵۶] .

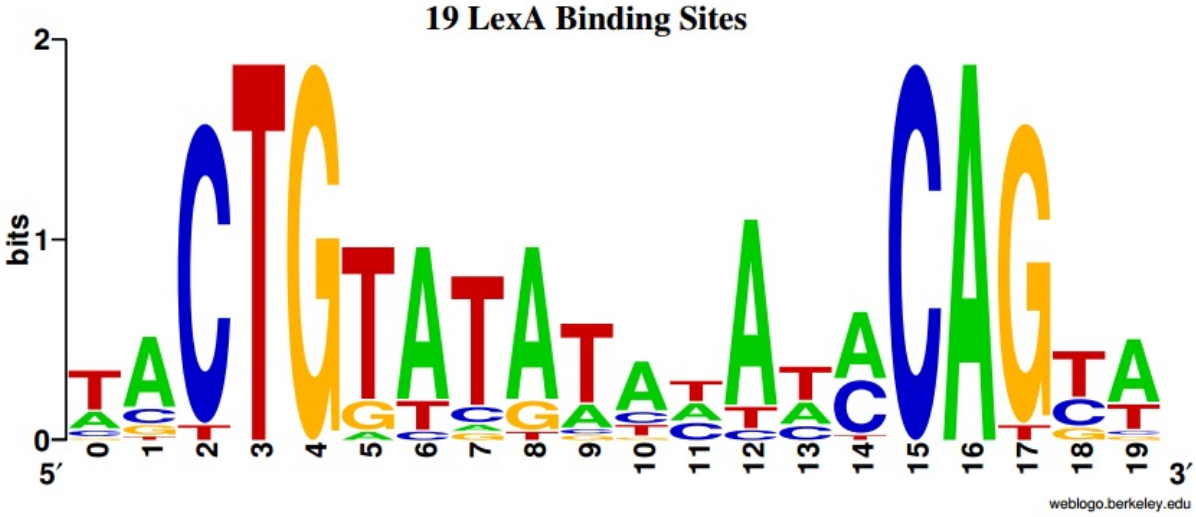
با توجه به این که، مسائل بهینه سازی قرن‌هاست توسط جامعه ریاضی مخصوصاً رشته تحقیق درعملیات مطالعه می‌شوند، بسیاری از فنون موفق این زمینه، در سال‌های گذشته، برای حل مسائل بهینه سازی «NP-سخت» بکار گرفته شده اند.

زیست‌شناسی مولکولی بدلیل این ویژگی که، ژنوم‌ها توالی‌یابی شده و داده‌های زیستی به وفور یافت می‌شوند وارد دوران پس از ژنومی شده است، با این حال بسیاری از فرآیندهای زیستی هنوز به خوبی درک نشده اند. یکی

از فرآیندها که مسئول تبدیل الگوی DNA [۱]شخص به یک واکنش پویا و تطبیق موجود زنده با نیازهای حیاتی می‌باشد بیان ژن است که به صورت انتخابی، ژن‌های روشن و خاموش را تغییر می‌دهد. تنها یک زیرمجموعه از ژن‌ها در هر سلول در مرحله رشد موجود زنده، حالت فیزیولوژیکی یا متابولیکی سلول و تحت هر مجموعه از شرایط محیطی فعال می‌باشند. بیان ژن فرآیندی است که در آن از اطلاعات درون ژن استفاده می‌شود تا یک محصول کاربردی به دست آید. محصول ژن‌ها عمدتاً پروتئین‌ها هستند و از محصولات غیر پروتئینی می‌توان به rRNA[۲] ، tRNA[۳]، snRNA [۴]اشاره کرد. مراحل مختلفی را می‌توان برای فرایند بیان ژن در نظر گرفت که عموماً شامل رونویسی، اتصال mRNA [۵]به ریبوزوم و ترجمه و تغییرات بعد از ترجمه یک پروتئین می‌باشد. درواقع، کدهای ژنتیکی که در رشته‌های DNA ذخیره شده‌اند، به وسیله بیان ژن تفسیر می‌شوند و خصوصیات و نحوه بیان ژن باعث به وجود آمدن خصوصیات قابل مشاهده در موجود زنده خواهد شد. به طور کلی ابتدا از روی ژن‌ها mRNA رونویسی شده و سپس به پروتئین ترجمه می‌شوند. رونویسی به طور معمول به وسیله پروتئین‌های خاص، با نام فاکتورهای رونویسی تسهیل می‌یابد که عملکردشان را به وسیله اتصال به قطعات DNA در مجاورت ژنی که در حال تنظیم شدن هستند، انجام می‌دهند [۴] . شناسایی چنین محل های اتصال DNA به نوبه خود مسأله مهمی است که به عنوان گامی اولیه و ضروری در درک تنظیم ژن محسوب می‌شود. در این پایان‌نامه، ما روی تکنیک‌های محاسباتی برای کشف محل‌های اتصال فاکتور رونویسی تمرکز می‌کنیم. فاکتورهای رونویسی که به عناصر توالی DNA متصل می‌شوند عملگر یا محل‌های اتصال تنظیمی هنگام شکل‌گیری کمپلکس های پروتئین- DNA نامیده شده‌اند. اکثر محل‌های اتصال، در نزدیکی محل شروع رونویسی، در بالادست آن نسبت به جهت رونویسی قرار گرفته‌اند. ناحیه DNA شامل این محل‌ها نواحی تنظیمی نامیده شده‌اند که به طور معمول برای پروکاریوت‌ها طولانی‌تر از ۱۰۰۰ نوکلئوتید نیست و کوتاه‌تر است. محل‌های اتصال تنظیمی معمولاً مخصوص پروتئین خاصی هستند که در توالی شبیه یکدیگرند.

[۱] Deoxyribonucleic acid
[۲] Transfer RNA
[۳] Small nuclear ribonucleic acid
[۴] Ribosomal ribonucleic acid
[۵] Messenger RNA

زمانی که محل‌های اتصال برای یک فاکتور رونویسی پروتئین با هم مرتب می‌شوند، الگوهایی به آسانی با نواحی بیشتر محافظت شده و کمتر محافظت شده، قابل مشاهده‌اند (شکل ۱.۱) که ما این الگوها را موتیف می‌نامیم. چون فاکتورهای رونویسی به طور معمول تعدادی از ژن‌های مختلف را تنظیم می‌کنند در مجموع به عنوان فاکتور تنظیم کننده رونویسی معرفی شده‌اند که موتیف یا امضا(اثر) اتصال می‌تواند در بین آنها تشخیص داده شود.



شکل۱.۱: لوگو توالی [۸۰] از ۱۹ جفت نوکلئوتید محل اتصال فاکتور رونویسی $LexA$ در $E.coli$

کشف تقریبی الگوهای تکرار شونده یا موتیف‌ها در توالی‌های زیستی، یک مسأله مهم و مورد مطالعه در زیست شناسی مولکولی محاسباتی است. مسأله موتیف‌یابی، اغلب در شناسایی سیگنال‌های تنظیمی درون توالی DNA و عناصر ساختاری و عملکردی درون توالی پروتئین ظاهر می‌شود. محل‌های اتصال واقعی نمونه‌های موتیف که ما می‌خواهیم در زمینه تنظیمی شناسایی کنیم، اغلب کوتاه (حداکثر ۳۰ نوکلئوتید با چند استثناء) و بدون فاصله هستند و از طول یکسانی برخوردار می‌باشند.

در این فصل به بیان تعاریف و مفاهیمی که در فصل‌های بعدی به آن‌ها نیاز می‌شود، می‌پردازیم.

۱.۱ مبانی زیست‌شناسی

DNA: با استفاده از ژنوم شرح کاملی از زندگی هر موجود زنده قابل بیان است که به طور خاص می‌تواند به این صورت مطرح شود: «توصیف مجموعه‌ای از دستورالعمل‌های دنبال شونده توسط موجود زنده به منظور رشد و نمو برای رسیدن به شکل نهایی آن».

ژنوم‌ها یک طرح اولیه از اشکال مختلف حیات است. زبان مورد استفاده توسط طبیعت برای کدگذاری حیات به وسیله کدهای DNA نشان داده می‌شود. یک ژنوم به صورت دئوکسی ریبونوکلئیک اسید (DNA) ساخته شده که در سلول‌های هر موجود زنده وجود دارد و متشکل از دو رشته محکم مارپیچ مانند از نوکلئوتیدها یا بازها است. چهار باز مختلف به نام های آدنین [۱] (A)، تیمین [۲](T)، سیتوزین [۳] (C) و گوانین [۴] (G) در DNA وجود دارند. ترتیب خاص از بازها توالی ژنومی نامیده می‌شود که مشخص کننده دستورالعمل‌های ژنتیکی دقیق برای ایجاد یک شکل خاص از زندگی است که ویژگی‌های منحصر به فرد خود را دارد.

دو رشته DNA در یک شکل مارپیچ معمولی به هم پیچ خورده‌اند و بازها در هر رشته توسط پیوندهای ضعیف به هم متصل شده‌اند که به اصطلاح جفت باز (bp) نامیده می‌شوند. دو رشته به صورت مکمل مقابل هم قرار دارند و تنها جفت شدن آدنین با تیمین ($A \leftrightarrow T$) و سیتوزین با گوانین ($C \leftrightarrow G$) مجاز است. در طول تقسیم سلولی، DNA قادر است خودش را تکثیر کند (همانند سازی). به این ترتیب پیوندهای بین جفت بازها در مولکول DNA از هم شکسته می‌شوند و سپس هر رشته به طور مستقیم رشته تکمیلی جدید را می‌سازد. براساس قانون مکمل، ژنوم جدید باید یک نسخه دقیق از قبلی باشد. البته، این فرآیند به طور کامل عاری از خطا نیست و تعدادی از بازها ممکن است از دست بروند، تکثیر شوند و یا به سادگی تغییر یابند. تغییرات در محتوای اصلی توالی DNA به عنوان جهش شناخته می‌شود. جهش‌ها در اغلب موارد مرگبار هستند و در سایر موارد می‌توان آن‌ها را کاملا بی ضرر و یا در دراز مدت پیشرو برای تکامل یک گونه تلقی کرد.

[۱] Adenine
[۲] Thymine
[۳] Cytosine
[۴] Guanine

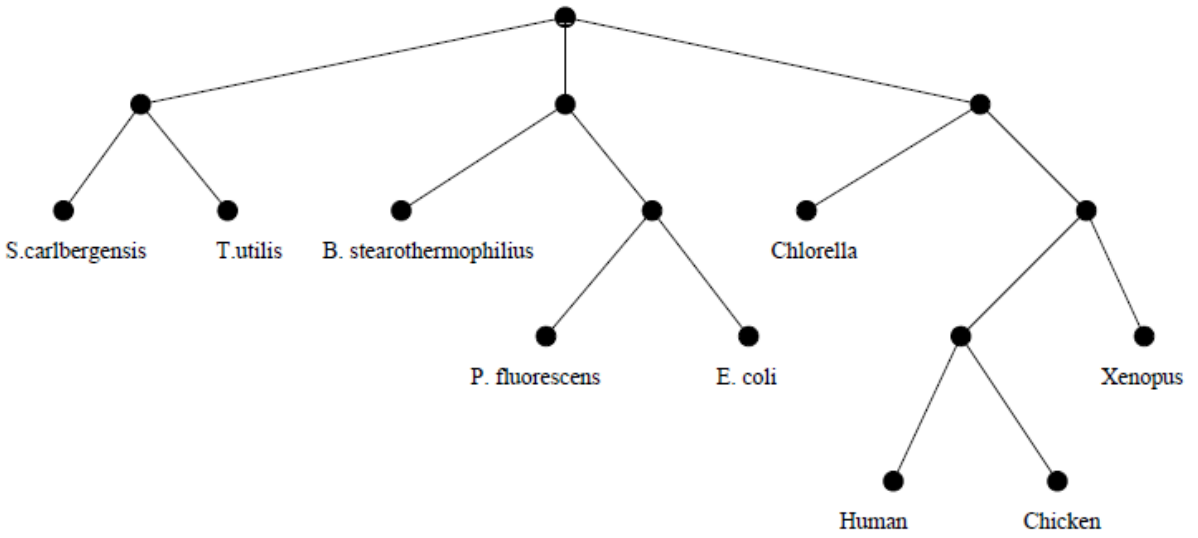
در انسان DNA در هر سلول متشکل از حدود ۳ میلیارد جفت باز است. فرآیند خواندن لیست مرتب شده بازها، از یک ناحیه ژنومی توالی یابی نامیده می‌شود. با توجه به محدودیت‌های موجود، نواحی توالی نمی‌تواند طولانی‌تر از چند صد جفت باز باشد. ساز و کار اساسی حیات در طبیعت برای تمام موجودات زنده مشابه است. در واقع تمام اطلاعات توصیف موجودات زنده، از طریق توالی‌های DNA در ژنوم آن، به وسیله ی یک کد جامع که به عنوان کد ژنتیکی شناخته می‌شود، کدگذاری شده است. این کد برای توصیف چگونگی ساخت پروتئین‌ها، که اجزای ساختاری از سلول‌ها و بافت‌ها، همچنین آنزیم‌های مورد نیاز ضروری برای واکنش‌های شیمیایی را فراهم می‌کنند، استفاده شده است. از همه توالی‌های DNA ،تنها بخش کوچکی از آن‌ها شامل اطلاعات کدگذاری(رمز گذاری) هستند(برای نمونه در انسان ها حدود ۱۰درصد از کل DNA رمزگذار است). اطلاعات نواحی رمزگذار DNA به صورت ژن‌ها سازمان‌دهی شده‌اند که در آن هر ژن مسئول رمزگذاری یک پروتئین متفاوت است [۵۶].

پروتئین: پروتئین یک مولکول بزرگ است که متشکل از چند آمینواسید است. در ژن کدکننده پروتئین، ترتیب آمینواسیدها را ترتیب نوکلئوتیدها مشخص می‌کنند. ۲۰ اسیدآمینه وجود دارد که هریک از آنها با کلمه سه حرفی از DNA مشخص می شوند. کلمات یا رشته های سه حرفی DNA رمز (codon) نامیده می‌شود. تناظر کدون‌ها و اسیدهای آمینه به عنوان "کد ژنتیکی" شناخته شده است. با توجه به اینکه $4^3 = 64$ کلمه سه حرفی از نوکلئوتیدها وجود دارد، اغلب اسیدهای آمینه با بیش از یک سه حرفی کدگذاری شده اند(آمینواسیدهای مترادف). برای مثال، آمینواسید پرولین(pro) توسط چهار سه تایی، CCT،CCC،CCA،CCG مشخص شده است. جایگزینی حتی یک آمینواسید در یک پروتئین، می‌تواند اثرات مرگباری داشته باشد (برای مثال بیماری مرگبار کم خونی سلول داسی شکل توسط تغییر یک آمینواسید در ژن کدکننده برای هموگلوبین ایجاد می‌شود). بنابراین ترادف‌های موجود در آمینو اسیدها، باعث کاهش احتمال جهش های تصادفی برخی نوکلئوتیدها، که می تواند اثرات مضر داشته باشد، می‌شود. بدین ترتیب برای ساخت یک پروتئین، DNA یک ژن، ابتدا به صورت تکRNA رشته‌ای پیامبر (mRNA) نسخه‌برداری می‌شود که این فرآیند را رونویسی می‌خوانند. RNA

(ریبونوکلئیک اسید) در واقع شبیه DNA است با این تفاوت که باز تیمین (T) با باز اوراسیل (U) جایگزین می‌شود. mRNA به عنوان الگو برای سنتز پروتئین عمل می‌کند به طوری که از هسته به سیتوپلاسم حرکت می‌کند. برخی از اجزای سلولی که ریبوزوم نامیده می‌شوند توالی سه حرفی mRNA را می‌خواند و آمینواسید مربوطه شناسایی می‌شود و اتصال آن‌ها منجر به توالی پروتئین می‌شود. روند ترجمه ژن به پروتئین نیازمند وجود سیگنال‌هایی برای شناسایی آغاز و پایان اطلاعات هر یک از نواحی کدگذاری است [۵۶].

علم ژنتیک : ژن‌ها واحد وراثت هستند که از پدر و مادر به فرزندان منتقل می‌شوند. هر ژن به تنهایی یا با ژن‌های دیگر مسئول یک یا چند ویژگی موجود زنده می‌باشد. هر ژن وظیفه‌ی مخصوصی در بدن دارد. این وظیفه شامل نقشه و دستورالعمل ساختن پروتئین‌ها در سلول است.

تکامل: شاخه ای از علم ژنتیک که به عنوان ژنتیک تکاملی شناخته شده و تغییرات ژنومی که در طول مدت زمان طولانی رخ داده است را بررسی می‌کند. برای تغییراتی که میزان اثرات آن زیاد باشد، زیرگونه‌های مختلف در داخل گونه افزایش می‌یابند و یا احتمالا گونه کاملا جدید را سبب می‌شوند. ساختار اصلی داده که برای نشان دادن روابط تکاملی موجود استفاده شده، درخت فیلوژنتیک است که در شکل ۲.۱ نمونه‌ای از آن را مشاهده می‌کنید.



شکل۲.۱: نمونه ای از درخت تکاملی [۵۶]

زمانی که دو ژنوم با هم مقایسه می‌شوند، هدف پیداکردن یک توالی از رویدادهای تکاملی است که با تأثیر روی ژنوم اول آن را به دومی تبدیل کرده است. هدف این است مجموعه ای صرفه جو از رویدادها و جوابی که یک کران پایین راجع به «میزان تکامل» بین دو ژنوم را فراهم می‌کند، پیدا شود. این کران، که به عنوان فاصله تکاملی تفسیر شده است، می‌تواند برای محاسبه درخت‌های فیلوژنتیک برای مجموعه‌ای از گونه‌ها استفاده شود. گونه‌هایی که در فاصله تکاملی کوتاه هستند باید در درخت، خیلی نزدیک به هم قرار داده شوند [۵۶].

باکتری اشریشیاکلی : باکتری «اشریشیا کلی» که معمولا با نام مخفف «ای کلی» یا E.coli نامیده می‌شود، باکتری گرم منفی میله‌ای شکل است که به طور شایع در بخش تحتانی روده حیوانات خون‌گرم یافت می‌شود و در سال ۱۸۵۵ توسط تئودور اشریخ[۱] کشف شد[۲۵]. درDNA حلقوی این باکتری باز گوانین (G) بیشتر از سایر بازها حضور دارد. اشریشیاکلی نخستین جانداری بود که با روش‌های مهندسی ژنتیک مورد دست‌ورزی

[۱] Theodor Escherich

ژن قرار گرفت. از اشریشیاکلی به عنوان یک موجود زنده مدل در پژوهش‌های ژنتیکی استفاده می‌شود. اولین توالی DNA کامل از یک ژنوم اشریشیاکلی در سال ۱۹۹۷ منتشر شده بود .

همردیفی چندگانه توالی: مقایسه توالی‌های ژنومی حاصل از افراد با گونه‌های متفاوت یا مشابه، یکی از اساسی‌ترین مسائل در زیست‌شناسی مولکولی است. این مقایسه با هدف شناسایی نواحی DNA با درجه حفاظت بالا(و بنابراین عملکردهای احتمالی مربوطه)، جهش‌های مرگبار نقطه‌ای، پیشنهاد روابط تکاملی، یا کمک به اصلاح خطاهای توالی‌یابی انجام می‌شوند.

یک توالی ژنومی می‌تواند به عنوان یک رشته بر الفبایی متشکل از ۴ کلمه نوکلئوتید یا ۲۰ کلمه برای شناسایی ۲۰ آمینواسید نشان داده شود. مرتب کردن یک مجموعه از توالی‌ها(یعنی، محاسبه یک همردیفی چندگانه توالی) عبارت است از ترتیب دادن آنها در یک ماتریس که هر سطر شامل یک توالی است که در صورت نیاز با قراردادن شکاف‌ها (با کاراکتر «-») در هر توالی به‌دست می‌آید، تا همه آنها هم طول شوند. در اینجا هدف شناسایی الگوهای مشترک است که به وسیله تلاش برای قراردادن کاراکترهای مشابه در هر ستون (تا آنجا که ممکن است) اتخاذ شده است. در زیر یک مثال ساده از همردیفی توالی‌های ATTCCGAC و ATCCTC و TTCCCTG آورده شده است. مثال نشان می‌دهد که TCC در بین همه توالی‌ها مشترک است.

```
A T T C C G A - C
- T T C C C - T G
A - T C C - - T C
```

مسأله همردیفی چندگانه توالی به عنوان یک مسأله بهینه‌سازی فرمول‌بندی شده است. معروف‌ترین تابع هدف برای همردیفی چندگانه، ایده‌هایش از دو توالی همردیف بهینه، تعمیم می‌یابند. این مسأله همردیفی دوگانه نامیده شده و به صورت زیر فرمول‌بندی می‌شود. هزینه‌های متقارن$\gamma(a,b)$ (یا به طور معادل برای سود) برای جایگزینی یک کلمه a با کلمه b و هزینه $\gamma(a,-)$ برای حذف یا وارد کردن کلمه a داده شده‌اند، کمینه هزینه (به