به نام خدا

تمرین سری اول
درس مقدمه ای بیوانفورماتیک
دکتر علی شریفی زارچی

فرزان رحمانی
۴۰۳۲۱۰۷۲۵

سوال اول

الف) **چگونه یک جهش در ژن می‌تواند باعث کاهش یا توقف بازسازی سلول‌های عضلانی شود؟**

بازسازی عضله به **سلول‌های ماهواره‌ای(satellite cells)** وابسته است که نوعی سلول‌های بنیادی در بافت عضله هستند و در پاسخ به آسیب فعال می‌شوند. این سلول‌ها تکثیر می‌شوند، به **میوبلاست‌ها** تمایز می‌یابند و در نهایت فیبرهای عضلانی جدیدی را برای ترمیم ناحیه آسیب‌دیده تشکیل می‌دهند. این فرآیند به شدت توسط **پروتئین‌ها و مسیرهای سیگنال‌دهی(proteins and signaling pathways)** مختلفی تنظیم می‌شود که توسط ژن‌های خاصی کدگذاری شده‌اند.

اگر یک جهش در ژنی رخ دهد که مسئول تولید پروتئین حیاتی برای بازسازی سلول‌های عضلانی باشد، این جهش می‌تواند به چندین روش فرآیند بازسازی را تحت تأثیر قرار دهد:

۱. **اختلال یا از دست دادن عملکرد پروتئین(Protein Dysfunction or Loss of Function):**

- جهش ممکن است منجر به **جهش جایگزینی** (missense mutation) شود، که در آن یک آمینواسید به یک آمینواسید دیگر تغییر می‌کند و عملکرد پروتئین تغییر می‌کند. همچنین ممکن است جهش به **جهش بی‌معنی** (nonsense mutation) منجر شود، که یک کدون توقف زودهنگام ایجاد می‌کند و منجر به تولید یک پروتئین ناقص و غیرعملکردی می‌شود.
- اگر پروتئین غیرعملکردی یا غایب باشد، فرآیندهای کلیدی مانند فعال‌سازی، تکثیر یا تمایز سلول‌های ماهواره‌ای ممکن است مختل شوند.

۲. **اختلال در مسیرهای سیگنال‌دهی (Disruption of Signaling Pathways):**

- بسیاری از پروتئین‌های دخیل در بازسازی عضله به عنوان **مولکول‌های سیگنال‌دهی(signaling molecules)** یا گیرنده‌ها عمل می‌کنند. اگر این پروتئین‌ها جهش پیدا کنند، **مسیرهای سیگنال‌دهی(signaling pathways)** که فعالیت سلول‌های ماهواره‌ای را تنظیم می‌کنند ممکن است مختل شوند و مانع از ترمیم مناسب عضله شوند.

۳. **اختلال در تکثیر یا تمایز سلول‌ها (Failure of Cell Proliferation or Differentiation):**

- جهش‌ها ممکن است ژن‌های کنترل‌کننده **چرخه سلولی(cell cycle)** را نیز تحت تأثیر قرار دهند و منجر به کاهش تکثیر سلول‌های ماهواره‌ای شوند. اگر سلول‌های ماهواره‌ای نتوانند به طور مؤثری تکثیر شوند، تعداد کمتری از سلول‌ها برای تمایز به فیبرهای عضلانی جدید در دسترس خواهد بود و بازسازی را مختل می‌کند.

ب) **وراثت جهش و اثرات وابسته به جنسیت**

از آنجایی که این جهش بر روی **کروموزوم X** قرار دارد، الگوی وراثت بیماری در مردان و زنان به دلیل ترکیب کروموزوم‌های جنسی آنها متفاوت خواهد بود:

- **مردان:(XY)**
  - مردان تنها **یک کروموزوم X** دارند. اگر آنها یک کروموزوم X حامل ژن جهش یافته را به ارث ببرند، **کروموزوم X دیگری** ندارند که نقص را جبران کند.
  - بنابراین، مردان **فنوتیپ بیماری(disease phenotype)** را بروز خواهند داد زیرا تنها نسخه موجود از ژن، جهش یافته است.
- **زنان:(XX)**
  - زنان **دو کروموزوم X** دارند. اگر یک کروموزوم X جهش یافته را به ارث ببرند، هنوز یک کروموزوم X دیگر دارند که ممکن است نسخه **سالم** ژن را حمل کند.
  - اگر جهش **مغلوب** باشد، زنان معمولاً **ناقل** خواهند بود و علائمی نشان نخواهند داد، زیرا آلل(allele) سالم روی کروموزوم X دیگر می‌تواند نقص را جبران کند.
  - با این حال، اگر جهش **غالب** باشد، حضور یک آلل(allele) جهش یافته برای بروز بیماری کافی است و بنابراین دختر نیز فنوتیپ بیماری را نشان می‌دهد.
  - علاوه بر این، **غیرفعال‌سازی کروموزوم X** (X-inactivation) می‌تواند نقش داشته باشد. در هر سلول زن، یکی از کروموزوم‌های X به طور تصادفی غیرفعال می‌شود. اگر کروموزوم X سالم در درصد بیشتری از سلول‌ها غیرفعال شود، دختر حتی اگر جهش مغلوب باشد، ممکن است علائم را نشان دهد.

خلاصه:

- **پسر** بیماری را بروز می‌دهد زیرا او تنها یک کروموزوم X دارد و این کروموزوم، جهش یافته است.
- **دختر** ممکن است ناقل باشد (اگر جهش مغلوب باشد) یا بیماری را بروز دهد (اگر جهش غالب باشد یا به دلیل غیرفعال‌سازی شدید کروموزوم X).

سوال دوم

الف) **تأثیر جهش در DNA پلیمراز بر توالی ژن و عملکرد پروتئین**

DNA پلیمراز نقش مهمی در تکثیر DNA دارد، زیرا مسئول سنتز رشته‌های جدید DNA با افزودن نوکلئوتیدهای مکمل به رشته قالب است. این آنزیم همچنین دارای توانایی **تصحیح خطا** است تا دقت بالایی را در طول تکثیر تضمین کند. اما جهشی در ژن کدکننده DNA پلیمراز که دقت آن را کاهش دهد، می‌تواند اثرات زیر را داشته باشد:

۱. **افزایش خطاهای تکثیر**:

- با کاهش دقت، DNA پلیمراز ممکن است **نوکلئوتیدهای نادرست**ی را اضافه کند (برای مثال، افزودن آدنین به جای سیتوزین). این اشتباهات می‌توانند منجر به **جهش‌های نقطه‌ای** شوند، مانند:
  - **جهش جایگزینی(Missense Mutation) :** تغییر یک نوکلئوتید که منجر به جایگزینی یک آمینواسید با آمینواسید دیگری در پروتئین می‌شود. این تغییر می‌تواند ساختار و عملکرد پروتئین را تغییر دهد و احتمالاً فعالیت یا پایداری آن را کاهش دهد.
  - **جهش بی‌معنی(Nonsense Mutation)** : تغییراتی که یک **کدون توقف زودهنگام** ایجاد می‌کند، که منجر به تولید پروتئینی ناقص و احتمالاً غیرعملکردی می‌شود.

- **جهش خاموش** :**(Silent Mutation)** تغییری که بر توالی آمینواسید پروتئین تأثیری ندارد، به دلیل وجود افزونگی در کد ژنتیکی. این جهش‌ها ممکن است تأثیر فوری بر عملکرد نداشته باشند، اما می‌توانند بیان ژن یا فرایند برش و اتصال (splicing) را تحت تأثیر قرار دهند.

۲. **جهش‌های جابجایی چارچوب(Frameshift Mutations)** :

- اگر DNA پلیمراز نتواند **افزودن یا حذف** نوکلئوتیدها (indels) را تصحیح کند، ممکن است جهش‌های جابجایی چارچوب ایجاد شود که **چارچوب خواندن** ژن را تغییر می‌دهد. این تغییر معمولاً منجر به تولید پروتئینی کاملاً متفاوت و غیرعملکردی می‌شود زیرا توالی آمینواسید پس از جهش تغییر می‌کند.

۳. **تأثیر بر عملکرد پروتئین**:

- جهش‌ها می‌توانند منجر به **پروتئین‌های تغییر یافته** شوند که ممکن است قابلیت عملکرد خود را از دست بدهند، کارایی کمتری داشته باشند، یا خواص مضری (مانند تاخوردگی نادرست یا تجمع) به دست آورند. این پروتئین‌های غیرعملکردی می‌توانند فرآیندهای سلولی را مختل کنند و ممکن است منجر به بیماری‌هایی مانند **سرطان** شوند، جایی که تجمع خطاهای تکثیر منجر به رشد غیرقابل کنترل سلولی می‌شود.

خلاصه :یک جهش در DNA پلیمراز که دقت آن را کاهش دهد، منجر به افزایش خطاهای تکثیر می‌شود که می‌تواند انواع مختلفی از جهش‌ها را در توالی‌های ژنی ایجاد کند. این جهش‌ها ممکن است منجر به تولید پروتئین‌های معیوب یا غیرعملکردی شوند که عملکرد سلولی را مختل کرده و منجر به بروز بیماری شوند.

ب) **اختلال در تکثیر DNA به دلیل جهش در DNA هلیکاز**

DNA هلیکاز برای تکثیر DNA ضروری است زیرا دو رشته‌ی DNA را باز می‌کند و آنها را از هم جدا می‌کند تا به عنوان قالب برای سنتز رشته‌های جدید استفاده شوند. اگر جهشی توانایی DNA هلیکاز را در باز کردن رشته‌های دوگانه DNA در نرخ طبیعی کاهش دهد، مشکلات زیر ممکن است به وجود آید:

۱. **توقف چنگال‌های تکثیر(Stalled Replication Forks)** :

- تکثیر DNA از سایت‌های خاصی به نام **مبداهای تکثیر(origins of replication)** آغاز می‌شود و به صورت دوطرفه ادامه می‌یابد و چنگال‌های تکثیر را تشکیل می‌دهد. اگر DNA هلیکاز نتواند DNA را به طور مؤثر باز کند، **چنگال‌های تکثیر** ممکن است متوقف شوند و فرآیند تکثیر کند یا حتی متوقف شود.

- توقف چنگال‌ها می‌تواند منجر به ایجاد **شکاف‌هایی** در رشته‌های تازه سنتز شده DNA شود که باعث ناقص ماندن تکثیر می‌شود.

۲. **افزایش تنش و سوپرپیچیدگی(Supercoiling)** :

- هنگامی که DNA هلیکاز DNA را باز می‌کند، **سوپرپیچیدگی** در جلوی چنگال تکثیر ایجاد می‌شود. اگر فعالیت هلیکاز کاهش یابد، یک **عدم تعادل** بین فرآیند باز کردن و سنتز توسط DNA پلیمراز ایجاد می‌شود.

- این می‌تواند باعث افزایش تنش و سوپرپیچیدگی در DNA شود که عبور دستگاه تکثیر را دشوارتر کرده و ممکن است منجر به **شکست‌های DNA** یا ناهنجاری‌های ساختاری شود.

۳. **تشکیل نواحی تک‌رشته‌ای:DNA**

- با کاهش فعالیت هلیکاز، برخی از نواحی DNA ممکن است به طور موقت **تک‌رشته‌ای** شوند اما برای مدت طولانی‌تری باز بمانند. این نواحی تک‌رشته‌ای بیشتر در معرض آسیب، مانند **عدم تطابق نوکلئوتیدها** یا تغییرات شیمیایی قرار دارند.

۴. **Replication Stress و ناپایداری ژنومی**:

- تکثیر ناقص یا کند می‌تواند منجر به **استرس تکثیر(Replication Stress)** شود، که وضعیتی است که احتمال آسیب به DNA و جهش‌ها را افزایش می‌دهد. با گذشت زمان، این مسئله می‌تواند منجر به **ناپایداری ژنومی** شود که یکی از ویژگی‌های بسیاری از بیماری‌ها، از جمله سرطان است.

خلاصه :یک جهش در DNA هلیکاز که توانایی آن را در باز کردن DNA دو رشته‌ای کاهش دهد، پیشرفت طبیعی تکثیر DNA را مختل می‌کند و منجر به توقف چنگال‌های تکثیر، افزایش تنش DNA و مشکلات ساختاری(stalled replication forks, increased DNA tension, and structural problems) می‌شود. این مسائل می‌توانند باعث تکثیر ناقص یا نادرست DNA شوند که به ناپایداری ژنومی و احتمال توسعه بیماری‌های مختلف کمک می‌کند.

سوال سوم

شناسایی و جداسازی **آنزیم‌های مقاوم به حرارت** (heat-resistant enzymes) از باکتری‌های گرمادوست (thermophilic bacteria) که در محیط‌های با دمای بالا مانند چشمه‌های آب گرم زندگی می‌کنند، می‌تواند برای کاربردهای صنعتی بسیار ارزشمند باشد. با این حال، در فرآیند **جداسازی و خالص‌سازی** (isolation and purification) این آنزیم‌ها، چالش‌های خاصی وجود دارد که باید برای حفظ پایداری آنها مد نظر قرار گیرد:

چالش‌های جداسازی پروتئین‌ها از باکتری‌های گرمادوست

۱. **حفظ پایداری پروتئین در دماهای پایین‌تر** :(Maintaining Protein Stability at Lower Temperatures)

- **آنزیم‌های مقاوم به حرارت** (heat-resistant enzymes) از باکتری‌های گرمادوست برای عملکرد بهینه در دماهای بالا (مثل ۷۰ تا ۱۰۰ درجه سانتی‌گراد) سازگار شده‌اند. هنگامی که این آنزیم‌ها در دماهای پایین‌تر (مثل دمای اتاق یا ٤ درجه سانتی‌گراد) جداسازی می‌شوند، ممکن است فعالیت خود را از دست بدهند یا ناپایدار شوند، زیرا یکپارچگی ساختاری آنها ممکن است خارج از محدوده دمای بهینه مختل شود.

۲. **مشکل در تخریب سلول** :(Cell Lysis Difficulty)

- باکتری‌های گرمادوست اغلب دارای دیواره‌های سلولی **قوی‌تر** (robust cell walls) برای تحمل شرایط سخت هستند. این مسئله فرآیند **تخریب سلول** (cell lysis) را نسبت به باکتری‌های مزوفیل (mesophilic) که در دماهای عادی زندگی می‌کنند، چالش‌برانگیزتر می‌کند. برای دسترسی به پروتئین‌های داخل سلولی بدون دناتوره کردن آنزیم‌ها، روش‌های تخریب سلول کارآمدتری مورد نیاز است.

۳. **هم‌خالص‌سازی سایر پروتئین‌های مقاوم به حرارت** :(Co-purification of Other Heat-Resistant Proteins)

- از آنجایی که بسیاری از پروتئین‌های موجود در باکتری‌های گرمادوست به حرارت مقاوم هستند، تمایز و جداسازی آنزیم مورد نظر دشوار می‌شود. روش‌های معمول دناتوره کردن با حرارت (heat-denaturation methods) که برای رسوب‌دهی پروتئین‌های غیرهدف در نمونه‌های مزوفیل استفاده می‌شوند، ممکن است در اینجا کارساز نباشند زیرا پروتئین‌های ناخواسته نیز ممکن است در دماهای بالا پایدار باقی بمانند.

۴. **آلودگی توسط پروتئازهای مقاوم به حرارت** :(Contamination by Heat-Stable Proteases)

- این باکتری‌ها ممکن است **پروتئازهای مقاوم به حرارت** (heat-stable proteases) تولید کنند که می‌توانند حتی در دماهای بالا، پروتئین‌ها را تخریب کنند. این پروتئازها ممکن است در طول جداسازی و خالص‌سازی با آنزیم مورد نظر تداخل کنند.

استراتژی‌های حفظ پایداری در طول جداسازی و خالص‌سازی

۱. **بهینه‌سازی شرایط دمایی** :(Optimizing Temperature Conditions)

- در طول جداسازی، بهتر است فرآیند در **دمای بالاتر** (مثل ۵۰ تا ۶۰ درجه سانتی‌گراد)، نزدیک به محدوده فعالیت بهینه آنزیم انجام شود. این کار می‌تواند به حفظ پایداری آنزیم کمک کند، به ویژه در مراحل اولیه استخراج.
- **محلول‌های بافر پایدارکننده حرارت** (heat stabilization buffers) نیز ممکن است استفاده شوند که به طور خاص برای حفظ ساختار پروتئین در دماهای بالا فرموله شده‌اند.

۲. **تنظیم pH و استفاده از عوامل پایدارکننده** :(Adjusting the pH and Using Stabilizing Agents)

- استفاده از بافرهایی با **pHبهینه** برای آنزیم می‌تواند از دناتوره شدن آن جلوگیری کند. علاوه بر این، افزودن **عوامل پایدارکننده** (stabilizing agents)مانند گلیسرول(glycerol) ، ترهالوز(trehalose) ، یا نمک‌های خاص (مثل آمونیوم سولفات) می‌تواند به حفظ پایداری پروتئین در طول خالص‌سازی کمک کند.
- **یون‌های فلزی** (metal ions)مانند Mg^{2+} یا Ca^{2+} نیز ممکن است در صورت نیاز برای حفظ پایداری و عملکرد آنزیم متالوآنزیمی (metalloenzyme) ضروری باشند.

۳. **روش‌های تخصصی تخریب سلول** :(Specialized Cell Lysis Methods)

- به جای استفاده از تکنیک‌های استاندارد تخریب سلول، **روش‌های مکانیکی** (mechanical methods)مانند **سونیکاسیون** (sonication)، **بید-بیتینگ** (bead-beating)، یا **هموژنیزاسیون تحت فشار بالا** (high-pressure homogenization)ممکن است در باز کردن دیواره‌های سخت سلولی باکتری‌های گرمادوست مؤثرتر باشند.
- درمان‌های آنزیمی با استفاده از **لیزوزیم** (lysozyme)یا آنزیم‌های لیز کننده خاص دیگر نیز ممکن است در صورت ترکیب با گرمایش ملایم برای بهبود کارایی تخریب سلول مفید باشند.

۴. **استفاده از گرمایش برای خالص‌سازی پروتئین** :(Heat Treatment for Protein Purification)

- با استفاده از مقاومت حرارتی آنزیم، می‌توان یک مرحله **گرمایش** (heat-treatment step)را به کار برد. با انکوباسیون لیزات سلولی در دمایی بالاتر از محدوده مقاومت آنزیم‌های مزوفیل (مثل ۶۰ تا ۸۰ درجه سانتی‌گراد)، پروتئین‌های ناخواسته ممکن است دناتوره و رسوب کنند و آنزیم مقاوم به حرارت در محلول باقی بماند.
- این مرحله می‌تواند به عنوان یک استراتژی **خالص‌سازی اولیه** (initial purification)برای کاهش حضور پروتئین‌های غیرمقاوم به حرارت استفاده شود.

۵. **استفاده از مهارکننده‌ها برای جلوگیری از پروتئولیز** :(Use of Inhibitors to Protect Against Proteolysis)

- افزودن **مهارکننده‌های پروتئاز** (protease inhibitors)که در دماهای بالا مؤثر هستند، می‌تواند از تخریب توسط پروتئازهای مقاوم به حرارت جلوگیری کند. از مهارکننده‌هایی مانند **PMSF** (phenylmethylsulfonyl fluoride)یا مهارکننده‌های مقاوم به حرارت تخصصی ممکن است استفاده شود.

۶. **تکنیک‌های کروماتوگرافی تمایلی** :(Affinity Chromatography Techniques)

- استفاده از **کروماتوگرافی تمایلی** (affinity chromatography)با تگ‌های خاص) مثل His-tag یا (GST-tag می‌تواند بسیار مؤثر باشد. آنزیم نشان‌دار شده می‌تواند به طور انتخابی بر روی رزین کروماتوگرافی جذب شود و یک فرآیند خالص‌سازی هدفمند و ملایم را فراهم کند که احتمال دناتوره شدن را کاهش می‌دهد.

۷. **پروتکل‌های خالص‌سازی سریع** :(Rapid Purification Protocols)

- هرچه آنزیم سریع‌تر خالص‌سازی شود، زمان کمتری را در شرایط بالقوه ناپایدار می‌گذراند. استفاده از **کروماتوگرافی سریع مایع پروتئین** (Fast Protein Liquid Chromatography - FPLC)یا دیگر روش‌های خالص‌سازی سریع زنجیره سرد می‌تواند زمان قرارگیری آنزیم در دماهای غیربهینه را کاهش دهد.

خلاصه:

چالش اصلی در جداسازی آنزیم‌های مقاوم به حرارت از باکتری‌های گرمادوست، حفظ پایداری آنزیم در طول فرآیند جداسازی و خالص‌سازی است. با بهینه‌سازی شرایط دمایی، استفاده از عوامل پایدارکننده، اعمال روش‌های تخصصی تخریب سلول، و استفاده از تکنیک‌های خالص‌سازی انتخابی، این آنزیم‌ها می‌توانند به طور مؤثر پایدار و خالص‌سازی شوند بدون این که فعالیت خود را از دست بدهند.

سوال چهارم

الف) **محاسبه Hamming Distance**

فاصله همینگ (Hamming Distance) معیاری برای اندازه‌گیری تفاوت ژنتیکی است که تعداد جایگاه‌هایی را محاسبه می‌کند که در آن نوکلئوتیدهای (nucleotides) متناظر در دو توالی **DNA** با طول برابر متفاوت هستند. این فاصله فقط برای توالی‌های با **طول برابر** قابل استفاده است و به درج (insertion) یا حذف (deletion) توجه نمی‌کند.

داریم:

- x=AGCTGAC
- y=AGCAGTC

برای یافتن فاصله همینگ، هر نوکلئوتید متناظر را مقایسه می‌کنیم:

Position	x	y	Different?
1	A	A	No
2	G	G	No
3	C	C	No
4	T	A	Yes
5	G	G	No
6	A	T	Yes
7	C	C	No

Hamming Distance = Number of differences = 2

فاصله همینگ (Hamming Distance) یعنی تعداد تفاوت‌ها که برابر با ۲ است.

پاسخ :فاصله همینگ بین توالی‌های x و y برابر با **۲** است.

ب) **محاسبه Edit Distance**

فاصله ویرایشی (Edit Distance) یا **فاصله لون‌اشتاین(Levenshtein Distance)** ، حداقل تعداد عملیات لازم برای تبدیل یک توالی به توالی دیگر را اندازه‌گیری می‌کند. عملیات مجاز عبارتند از:

- **درج (Insertion)** یک کاراکتر
- **حذف (Deletion)** یک کاراکتر
- **جایگزینی (Substitution)** یک کاراکتر با کاراکتر دیگر

فاصله ویرایشی اطلاعات بیشتری درباره تفاوت ژنتیکی ارائه می‌دهد زیرا به **درج‌ها (insertions)** و **حذف‌ها (deletions)** توجه می‌کند، نه فقط جایگزینی‌ها (substitutions). بنابراین، برای توالی‌هایی با **طول‌های متفاوت** انعطاف‌پذیرتر است.

برای محاسبه فاصله ویرایشی بین x و y، از **برنامه‌نویسی پویا (dynamic programming)** استفاده می‌کنیم. ماتریسی به نام D ایجاد می‌کنیم که در آن D[i][j] فاصله ویرایشی بین اولین i کاراکتر از x و اولین j کاراکتر از y را نشان می‌دهد.

		A	**G**	**C**	**A**	**G**	**T**	**C**
	0	1	2	3	4	5	6	7
A	1	0	1	2	3	4	5	6
G	2	1	0	1	2	3	4	5
C	3	2	1	0	1	2	3	4
T	4	3	2	1	1	2	3	4
G	5	4	3	2	2	1	2	3
A	6	5	4	3	2	2	2	3
C	7	6	5	4	3	3	3	2

مقدار موجود در D[7][7] برابر **۲** است که حداقل تعداد ویرایش‌های لازم را نشان می‌دهد.

پس فاصله ویرایشی بین توالی‌های x و y برابر با **۲** است.

مقایسه: فاصله ویرایشی (Edit Distance) و فاصله همینگ(Hamming Distance)

- **فاصله همینگ (Hamming Distance)** فقط به **جایگزینی‌ها (substitutions)** توجه می‌کند و نیازمند توالی‌هایی با طول برابر است. این روش محدود است زیرا به درج‌ها و حذف‌ها توجه نمی‌کند.
- **فاصله ویرایشی (Edit Distance)** جامع‌تر است زیرا به **درج‌ها(insertions)** ، **حذف‌ها (deletions)** و **جایگزینی‌ها (substitutions)**توجه می‌کند. این مسئله باعث می‌شود که برای توالی‌هایی با طول‌های متفاوت یا در مواردی که جهش‌ها شامل تغییرات ساختاری مانند درج یا حذف (indels) هستند، کاربرد بیشتری داشته باشد.

نتیجه‌گیری :در حالی که هر دو فاصله در این مورد برابر با ۲ بودند، فاصله ویرایشی نمای وسیع‌تری از تغییرات ژنتیکی ارائه می‌دهد، زیرا می‌تواند تفاوت‌های ناشی از درج‌ها و حذف‌ها را که در تغییرات تکاملی (evolutionary changes) رایج هستند، شناسایی کند.

سوال پنجم

تراز کردن توالی‌های ($x = \text{GCAC}$) و ($y = \text{GCC}$) با استفاده از سیستم امتیازدهی داده شده:

- Match: +2
- Mismatch: -2
- Gap open: -5
- Gap extend: -1

ما از روش برنامه‌ریزی پویا برای پر کردن ماتریس تراز و محاسبه امتیاز تراز بهینه استفاده می‌کنیم.

مرحله ۱: مقداردهی اولیه ماتریس‌ها

ما سه ماتریس خواهیم داشت:

۱. ماتریس امتیاز (M): برای امتیاز تراز بهینه(optimal alignment).

۲. ماتریس gap در x (Ix): برای ترازهایی که فاصله‌ای در (x) ایجاد می‌شود.

۳. ماتریس فاصله در y (Iy): برای ترازهایی که فاصله‌ای در (y) ایجاد می‌شود.

فرض کنید ($M(i,j)$)، ($Ix(i,j)$) و ($Iy(i,j)$) به ترتیب امتیاز در موقعیت ((i,j)) در ماتریس‌ها هستند.

مرحله ۲: مقداردهی اولیه ماتریس‌ها

$$M(0,0) = 0$$

$$Ix(i,0) = M(i,0) = -5 - (i-1) \times 1 \; for \; (i \geq 1)$$

$$Iy(0,j) = M(0,j) = -5 - (j-1) \times 1 \; for \; (j \geq 1)$$

مرحله ۳: پر کردن ماتریس‌ها

$For \; (i \; \geq 1) \; and \; (j \; \geq 1)$:

- $Ix(i,j) = \max(Ix(i-1,j) - 1, M(i-1,j) - 5)$
- $Iy(i,j) = \max(Iy(i,j-1) - 1, M(i,j-1) - 5)$
- $M(i,j) = \max\left(M(i-1,j-1) + s(x_i, y_j), Ix(i,j), Iy(i,j)\right), where \; s(x_i, y_j) \; is$:
  - $+2 \; if \; x_i = y_j \; (match)$
  - $-2 \; if \; x_i \neq y_j \; (mismatch)$

ماتریس M:

		G	C	C
	0	-5	-6	-7
G	-5	2	-3	-4
C	-6	-3	4	-1
A	-7	-4	-1	2
C	-8	-5	-2	1

کد بالا به شکل زیر می شود:

```
import numpy as np
x = "GCAC"
```

```
y = "GCC"
match_score = 2
mismatch_score = -2
gap_open = -5
gap_extend = -1
m, n = len(x), len(y)
M = np.zeros((m + 1, n + 1))
Ix = np.zeros((m + 1, n + 1))
Iy = np.zeros((m + 1, n + 1))
M.fill(-np.inf)
Ix.fill(-np.inf)
Iy.fill(-np.inf)
M[0, 0] = 0
for i in range(1, m + 1):
    Ix[i, 0] = gap_open + (i - 1) * gap_extend
    M[i, 0] = gap_open + (i - 1) * gap_extend
for j in range(1, n + 1):
    Iy[0, j] = gap_open + (j - 1) * gap_extend
    M[0, j] = gap_open + (j - 1) * gap_extend
for i in range(1, m + 1):
    for j in range(1, n + 1):
        if x[i - 1] == y[j - 1]:
            score = match_score
        else:
            score = mismatch_score
        Ix[i, j] = max(Ix[i - 1, j] + gap_extend, M[i - 1, j] + gap_open)
        Iy[i, j] = max(Iy[i, j - 1] + gap_extend, M[i, j - 1] + gap_open)
        M[i, j] = max(M[i - 1, j - 1] + score, Ix[i, j], Iy[i, j])
alignment_score = M[m, n]
```

مرحله ٤: بازگشت به عقب (Traceback) برای هم‌ترازی بهینه

امتیاز بهینه برابر 1+ در D[4][3] است. حالا مراحل traceback را انجام می‌دهیم تا به هم ترازی بهینه برسیم.

نتیجه:

امتیاز تراز بهینه برای توالی‌های $(x = \text{GCAC})$ و $(y = \text{GCC})$ برابر با 1+ است. همچنین هم ترازی بهینه در ادامه آمده است:

$$G\ C\ A\ C$$

$$G\ C\ -\ C$$

این امتیاز نشان‌دهنده بهترین تراز ممکن با در نظر گرفتن جریمه‌های opening and extending gaps، و همچنین matches and mismatches می‌باشد.

سوال ششم

الف) ایده اصلی MUSCLE (Multiple Sequence Comparison by Log-Expectation) انجام هم ترازی چند توالی با دقت و سرعت بالا است. از ترکیبی از تکنیک‌ها، از جمله تخمین فاصله سریع با استفاده از شمارش k-mer، هم‌ترازی پیشرونده با امتیازدهی پروفایل (نمره ورود به سیستم انتظار) و اصلاح با استفاده از پارتیشن‌بندی محدود وابسته به درخت استفاده می‌کند(fast distance estimation using k-mer counting, progressive alignment with profile scoring (log-expectation score), and refinement using tree-dependent restricted partitioning). هدف MUSCLE ارائه هم‌ترازی‌های با کیفیت بالا با بهبود مکرر هم‌ترازی اولیه از طریق این مراحل است که آن را برای مجموعه داده‌های کوچک و بزرگ مناسب می‌سازد.

ب) در زمینه این مقاله، "hits" به ترازهای توالی با امتیاز بالا اشاره دارد که در طول جستجوهای پایگاه داده به دست می آیند. به عنوان مثال، هنگام جستجوی یک دنباله پروتئین در برابر یک پایگاه داده با استفاده از ابزارهایی مانند PSI-BLAST،"hits" دنباله هایی در پایگاه داده هستند که به خوبی با دنباله پرس و جو همسو(query sequence) می شوند و معمولاً یک آستانه شباهت مشخص را برآورده می کنند (مثلاً دارای یک مقدار e-value کمتر از ۰/۰۱).

ج) MUSCLE از سه مرحله اصلی تشکیل شده است:

۱. Draft Progressive Stage:

هدف: به سرعت یک تراز چند توالی اولیه ایجاد کنید.

فرآیند:

- فواصل k-mer جفتی را برای ایجاد یک درخت راهنمای اولیه با استفاده از روش خوشه‌بندی UPGMA محاسبه می‌کند.
- یک تراز را به تدریج بر اساس این درخت می سازد.

Improvement: تقریب سریع تراز را فراهم می کند و به عنوان نقطه شروع برای اصلاح عمل می کند.

۲. Improved Progressive Stage:

هدف: افزایش دقت تراز اولیه.

فرآیند:

- فاصله ها را با استفاده از فاصله Kimura بر اساس تراز اولیه دوباره محاسبه می کند.
- یک درخت راهنمای جدید می سازد و توالی ها را مطابق با این درخت دقیق تر تراز می کند.

Improvement: خطاهای احتمالی را از محاسبه فاصله تقریبی k-mer در مرحله اول تصحیح می کند و ساختار تراز را اصلاح می کند.

۳. Refinement Stage:

هدف: بهبود بیشتر دقت هم ترازی با به حداقل رساندن ناهماهنگی ها.

فرآیند:

- به طور مکرر یک لبه را از درخت راهنما انتخاب می کند، درخت را به زیردرختان تقسیم می کند و دنباله های این زیردرخت ها را دوباره تراز می کند.
- تراز جدید را در صورتی می پذیرد که امتیاز مجموع جفت ها (SP) را بهبود بخشد.

بهبود: به طور مکرر تراز را بهینه می کند، خطاهای وارد شده در مراحل پیش رونده را کاهش می دهد و به دقت تراز بیشتر می رسد.

د) روش k-mer شباهت زوجی دنباله ها(**pairwise similarity** of sequences) را با شمارش دنباله های فرعی مشترک (k-mers) با طول ثابت k محاسبه می کند. توالی‌های مرتبط معمولاً k-mer مشترک بیشتری نسبت به توالی‌های غیرمرتبط دارند. این روش از نیاز به هم‌ترازی کامل توالی اجتناب می‌کند و امکان تخمین سریع فواصل بین دنباله‌ها را بر اساس کسر k-mers مشترک فراهم می‌کند.

فاصله k-mer به خوبی با شباهت توالی ارتباط دارد و راهی سریع و کارآمد برای تقریب روابط توالی بدون انجام یک تراز کامل ارائه می دهد.

هـ) در مرحله پالایش، MUSCLE با استفاده از پارتیشن بندی محدود وابسته به درخت(ee-dependent restricted partitioning)، هم ترازی را بهبود می بخشد:

- یک لبه در درخت راهنما انتخاب می کند و درخت را به دو درخت فرعی تقسیم می کند.
- دنباله‌های هر زیردرخت دوباره تراز می‌شوند و با هم‌تراز کردن این زیرشاخه‌ها یک هم‌ترازی چندگانه جدید ایجاد می‌شود.
- اگر این تراز جدید امتیاز کلی مجموع جفت ها (SP) را بهبود بخشد، حفظ می شود. در غیر این صورت دور انداخته می شود.

این مرحله بسیار مهم است زیرا ناهماهنگی های محلی معرفی شده در مراحل پیش رونده را اصلاح می کند. این تضمین می کند که هم ترازی نهایی بیش از حد به دقت درخت راهنمای اولیه وابسته نیست و امکان تنظیم دقیق تراز توالی را فراهم می کند که منجر به نمایش دقیق تری از روابط تکاملی بین دنباله ها می شود.

و) MUSCLE به دلیل ویژگی های زیر برای مجموعه داده های بزرگ مناسب است:

- کارایی روش k-mer: استفاده از شمارش k-mer برای تخمین فاصله اولیه بسیار سریع است و زمان محاسباتی را به طور قابل توجهی در مقایسه با روش های تراز کامل کاهش می دهد.
- Progressive alignment strategy: MUSCLE با ایجاد تراز بر اساس درختان راهنما، از هزینه محاسباتی تراز کردن همه دنباله ها به طور همزمان جلوگیری می کند.
- Iterative refinement: در حالی که مرحله پالایش از نظر محاسباتی فشرده‌تر است، می‌توان آن را حذف کرد (با استفاده از MUSCLE-p) برای نتایج سریع‌تر در صورت نیاز، و آن را برای مجموعه‌های داده بسیار بزرگ قابل تطبیق می‌کند.
- مقیاس پذیری(Scalability): پیچیدگی زمانی MUSCLE (with the option to skip refinement) با افزایش تعداد دنباله ها به خوبی مقیاس می شود و به آن اجازه می دهد هزاران دنباله را به طور موثر بر روی سخت افزار محاسباتی استاندارد مدیریت کند.

این ویژگی‌ها با هم ترکیب می‌شوند تا MUSCLE را سریع و دقیق بسازند، و آن را به یک انتخاب ترجیحی برای کارهای هم‌ترازی چند توالی، به‌ویژه هنگام کار با مجموعه داده‌های بزرگ تبدیل می‌کنند.

پایان