

به نام خدا



دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)

درس پردازش دادگان حجیم

بهبود کارایی الگوریتم BWA در بستر GPU

استاد : دکتر محمدرضا پورفرد

سرپرست تیم تدریسیاری: دکتر مریم طالبی

استادیار طراح تمرین: امیرمحمد تلیسچیان تبریزی

تمرین شماره سوم

پاییز ۱۴۰۴

هدف تمرین

هدف از این تمرین، آشنایی عملی دانشجویان با تحلیل و بهینه‌سازی کدهای پردازش داده‌های حجیم در یک مسئله واقعی از حوزه بیوانفورماتیک است. در این تمرین، دانشجویان با بررسی یک پیاده‌سازی موجود از الگوریتم BWA، نحوه مدیریت حافظه و اجرای موازی در یک نرم‌افزار واقعی را تحلیل کرده و با تمرکز بر بهبود استفاده از حافظه اشتراکی (Shared Memory)، عملکرد برنامه را از نظر زمان اجرا، مصرف حافظه و مقیاس‌پذیری ارتقا می‌دهند. همچنین اجرای الگوریتم بر روی داده‌های Whole Exome Sequencing به دانشجویان کمک می‌کند تا چالش‌های عملی پردازش داده‌های بزرگ در مقیاس واقعی را تجربه کنند.

پیش‌نیازها

- تسلط نسبی بر زبان برنامه‌نویسی C/C++ و توانایی خواندن و تحلیل کدهای موجود
- آشنایی با مفاهیم برنامه‌نویسی موازی و حافظه اشتراکی (Shared Memory)، از جمله threadها و همگام‌سازی
- آشنایی کلی با مفاهیم پایه بیوانفورماتیک، به‌ویژه هم‌ترازی توالی‌های ژنومی

تمرین سوم

الگوریتم **BWA (Burrows-Wheeler Aligner)** یکی از پرکاربردترین الگوریتم‌ها برای هم‌ترازی (Alignment) توالی‌های کوتاه DNA با یک ژنوم مرجع است. این الگوریتم بر پایه **Burrows-Wheeler Transform (BWT)** و ساختار داده **FM-Index** طراحی شده و امکان جست‌وجوی سریع و کم‌مصرف توالی‌ها را در داده‌های ژنومی بسیار بزرگ فراهم می‌کند.

BWA با استفاده از ایندکس فشرده ژنوم مرجع، توالی‌های ورودی (Reads) را به‌صورت کارآمد روی ژنوم نگاشت می‌کند و با پشتیبانی از خطاهای توالی‌یابی مانند mismatch و gap، دقت بالایی در هم‌ترازی ارائه می‌دهد. به دلیل سرعت بالا و مصرف حافظه نسبتاً کم، این الگوریتم به‌طور گسترده در پردازش داده‌های **Whole Genome Sequencing** و **Whole Exome Sequencing** مورد استفاده قرار می‌گیرد.

Whole Genome Sequencing (WGS) به فرآیند توالی‌یابی کل ژنوم یک موجود زنده اشاره دارد. در این روش، تمامی نواحی ژنومی شامل نواحی کدکننده (Exons)، غیرکدکننده، اینترون‌ها، نواحی تنظیمی و بین‌ژنی توالی‌یابی می‌شوند و جامع‌ترین دید را از اطلاعات ژنتیکی فراهم می‌کند، اما به دلیل حجم بسیار بالای داده‌ها، نیازمند منابع محاسباتی و حافظه زیادی است.

در مقابل، **Whole Exome Sequencing (WES)** تنها بر روی نواحی کدکننده پروتئین (Exons) تمرکز دارد. این نواحی حدود ۱ تا ۲ درصد از کل ژنوم انسان را تشکیل می‌دهند، اما بخش عمده‌ای از جهش‌های مؤثر بر بیماری‌ها در همین نواحی رخ می‌دهد. به همین دلیل، **WES** با حجم داده کمتر، هزینه پایین‌تر و سرعت پردازش بالاتر، گزینه‌ای رایج در مطالعات ژنتیکی و بالینی محسوب می‌شود.

در این تمرین، استفاده از داده‌های **Whole Exome Sequencing** این امکان را فراهم می‌کند تا دانشجویان ضمن کار با داده‌های واقعی و نسبتاً بزرگ، چالش‌های پردازش و بهینه‌سازی الگوریتم‌ها را بدون پیچیدگی محاسباتی بسیار بالای **WGS** تجربه کنند.

در این تمرین، دانشجویان با بررسی پیاده‌سازی **BWA** و بهبود آن از نظر مدیریت حافظه اشتراکی، با چالش‌های عملی بهینه‌سازی الگوریتم‌های بیگ‌دیتا در کاربردهای واقعی آشنا خواهند شد.

بخش اول : بررسی و بهبود پیاده سازی الگوریتم BWA

در این بخش، یک پیاده سازی از الگوریتم BWA در اختیار دانشجویان قرار می گیرد که در مخزن گیت هاب زیر قابل دسترسی است:

<https://github.com/minhhpham/bwa>

دانشجویان موظفاند کد منبع ارائه شده را به صورت کامل بررسی کرده و ساختار، منطق اجرا و نحوه مدیریت حافظه در آن را تحلیل کنند. تمرکز اصلی این بخش بر شناسایی و بهبود نحوه استفاده از حافظه اشتراکی (Shared Memory) در پیاده سازی موجود است.

در این بخش لازم است شما :

۱. ساختار کلی کد و اجزای اصلی آن را تحلیل کرده و بخش های مرتبط با پردازش موازی و مدیریت حافظه را شناسایی نمایند.
۲. نحوه استفاده فعلی از حافظه (به ویژه داده های مشترک بین thread ها یا پردازش ها) را بررسی کرده و نقاط ضعف یا ناکارآمدی احتمالی را توضیح دهند.
۳. پیشنهادهای بهبود برای استفاده بهینه تر از Shared Memory ارائه داده و این بهبودها را در کد اعمال کنند.
۴. نسخه اصلاح شده کد را به گونه ای پیاده سازی کنند که منجر به بهبود عملکرد از نظر مصرف حافظه و زمان اجرا شود، بدون آنکه صحت نتایج الگوریتم تغییر کند.

در این بخش شما باید از داده های whole exome که در اختیارتان قرار میگیرد استفاده کنید و برنامه خود را تست کنید.

بخش دوم (امتیازی) :

در این بخش، دانشجویان باید نسخه نهایی پیاده سازی الگوریتم BWA (پس از اعمال بهبودهای بخش اول) را بر روی داده های Whole Genome Sequencing (WGS) اجرا و ارزیابی کنند. هدف این بخش، بررسی رفتار و مقیاس پذیری الگوریتم در مواجهه با داده هایی با حجم بسیار بزرگ تر و تحلیل چالش های عملی پردازش داده های ژنومی در مقیاس کل ژنوم است.

در این بخش باید :

۱. پیاده سازی بهبود یافته الگوریتم BWA را بر روی داده های WGS اجرا کنند.
۲. توجه داشته باشند که با افزایش حجم داده ها، مشکلاتی مانند race condition و ناسازگاری های احتمالی در حافظه اشتراکی ممکن است رخ دهد. دانشجویان باید این موارد را شناسایی کرده، نحوه

برخورد الگوریتم با آن‌ها را تحلیل کنند و در صورت امکان راهکارهایی برای کاهش این مشکلات ارائه دهند.

۳. نتایج به‌دست‌آمده از اجرای الگوریتم بر روی داده‌های Whole Genome را با نتایج حاصل از داده‌های Whole Exome (بخش اول) مقایسه کرده و تفاوت‌ها و چالش‌های مشاهده‌شده را تحلیل کنند.

بخش سوم (امتیازی) : طراحی معماری سیستم پردازش پیوسته داده‌های ژنومی

در این بخش، دانشجویان باید معماری یک سیستم بهینه برای پردازش پیوسته داده‌های ژنومی را طراحی و شرح دهند. در این سیستم، داده‌های ژنومی به صورت جریان داده (stream) وارد الگوریتم BWA می‌شوند و نتایج پس از پردازش، به‌طور پیوسته آماده خواهند شد. به عبارت دیگر، داده‌ها در یک صف (queue) قرار گرفته و به ترتیب پشت سر هم به الگوریتم منتقل می‌شوند.

در این بخش شما باید مراحل زیر را طی کنید :

۱. طراحی کلی سیستم

- معماری پیشنهادی شامل بخش‌های دریافت داده، صف‌بندی، مدیریت پردازش و خروجی نتایج را شرح دهند.
- نقش هر بخش در پردازش داده‌های حجیم و زمان واقعی (real-time) مشخص شود.

۲. بهینه‌سازی پردازش موازی و حافظه

- نحوه مدیریت Shared Memory و همگام‌سازی بین پردازش‌ها توضیح داده شود.
- راهکارهایی برای کاهش race condition و افزایش کارایی سیستم ارائه گردد.
- پیشنهادات برای افزایش throughput و کاهش latency سیستم ارائه شود.

۳. مقیاس‌پذیری و انعطاف‌پذیری

- سیستم باید بتواند با افزایش حجم داده یا افزایش تعداد Reads به صورت پیوسته، بدون کاهش قابل توجه کارایی، مقیاس‌پذیر باشد.
- در صورت امکان، توضیح داده شود که چگونه می‌توان سیستم را به پردازش‌های توزیع‌شده یا کلاسترهای محاسباتی گسترش داد.

۴. مستندسازی معماری

- دانشجویان باید یک نمودار معماری (Block Diagram) از سیستم ارائه دهند که جریان داده‌ها، اجزای اصلی، نقاط پردازش موازی و مدیریت صف‌ها را نشان دهد.
- توضیح متناظر برای هر بخش و تعامل بین اجزا نوشته شود.

تحويل تمرین :

۱. همه فایل‌ها و گزارش‌ها باید به صورت منظم، مستند و قابل اجرا باشند.
۲. گزارش‌ها می‌توانند به صورت PDF یا Word ارائه شوند و کدها در قالب فایل‌های اصلی پروژه باشند.
۳. نتایج آزمایش‌ها و نمودارها باید شامل مقایسه عملکرد قبل و بعد از بهبود و تحلیل دقیق باشند.

پیوست

کد ارائه شده در تمرین، بر اساس پیاده‌سازی الگوریتم BWA است و روش اجرای آن مشابه پروژه BWA-MEM می‌باشد. برای اجرای صحیح برنامه و آزمایش روی داده‌های ژنومی، دانشجویان می‌توانند از README پروژه مرجع استفاده کنند.

<https://github.com/lh3/bwa>

آدرس پروژه BWA-MEM :

برای اجرای کد ارائه‌شده در تمرین، بهتر است ابتدا README پروژه مرجع BWA-MEM را مطالعه کنید. این فایل شامل مراحل نصب، کامپایل و اجرای برنامه است و مثال‌هایی از دستورات پایه برای هم‌ترازی داده‌ها ارائه می‌دهد که می‌تواند برای اجرای تمرین بسیار مفید باشد.

پس از مطالعه README، محیط اجرای برنامه باید آماده شود. برای این منظور، پیش‌نیازهای نرم‌افزاری مانند gcc، make و ابزارهای خط فرمان لینوکس نصب شده و کد با استفاده از دستورات ارائه‌شده در README کامپایل می‌شود. همچنین بررسی مسیر فایل‌ها و داده‌های نمونه ضروری است تا از اجرای صحیح برنامه اطمینان حاصل شود.

برای اجرای برنامه، باید کد را با پارامترهایی مشابه BWA-MEM و داده‌های ورودی مورد نظر Whole Exome یا Whole Genome اجرا کنید. توجه به حجم داده‌ها و تخصیص منابع مناسب حافظه و پردازش موازی اهمیت بالایی دارد. علاوه بر این، ثبت نتایج خروجی و زمان اجرا برای مقایسه قبل و بعد از اعمال بهبودها ضروری است.

برای دریافت داده‌های whole exome می‌توانید از آدرس زیر استفاده کنید :

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR183/ERR183375/ERR183375_1.fastq.gz

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR183/ERR183375/ERR183375_2.fastq.gz

همچنین یکی از ورودی‌های دستوراتی که اجرا می‌کنید شامل ژنوم رفرنس می‌باشد که عملیات نگاشت با استفاده از آن انجام می‌شود. برای دریافت آن می‌توانید از آدرس زیر استفاده کنید :

ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg38/Homo_sapiens_assembly38.dict

ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg38/Homo_sapiens_assembly38.fasta.64.alt

ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg38/Homo_sapiens_assembly38.fasta.fai

ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg38/Homo_sapiens_assembly38.fasta.gz

این فایل‌ها را با استفاده از دستور wget در ترمینال لینوکس می‌توانید دانلود کنید.