دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده ریاضی و علوم کامپیوتر

پروژه درس مبانی بیوانفورماتیک

Paired de Bruijn Graphs: A Novel Approach for Incorporating Mate Pair Information into Genome Assemblers

نگارش

کیارش مختاری دیزجی - ۹۸۳۰۰۳۲

استاد

دکتر فاطمه زارع میرک آباد

بهمن ۱۴۰۲

فهرست مطالب

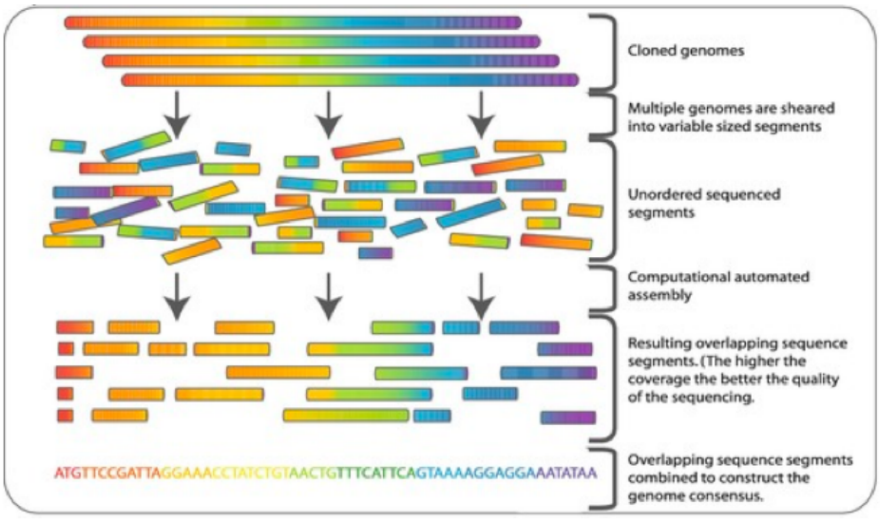
عنوان صفحه

فصل اول

مقدمه

۱-۱ مسئله زیستی چیست؟

مشکل زیست‌شناختی که در این گزارش سعی در پاسخ‌گویی به آن را داریم، مونتاژ ژنوم است. هدف بازسازی توالی اصلی یک ژنوم از مجموعه‌ای از توالی‌های کوتاه DNA (read) است، که به دلیل طبیعت تکراری DNA و وجود خطا در فرآیند توالی‌یابی، یک پازل پیچیده محسوب می‌شود.[۱]



شکل ۱-۱: مونتاژ ژنوم[۲]

بنابراین ورودی مسئله یک توالی از readها می باشد که هدف مونتاژ کردن ژنوم است.

۱-۲ چالش از دیدگاه آزمایشگاهی

در آزمایشگاه، تکنیک استفاده شده برای جمع‌آوری داده‌ها برای مونتاژ ژنوم، توالی‌یابی DNA است. به‌طور خاص، تکنولوژی‌های توالی‌یابی نسل جدید (NGS) قادر به تولید تعداد زیادی readهای کوتاه DNA از یک ژنوم به صورت سریع هستند.

تولید کامل ژنوم در آزمایشگاه به دلیل پیچیدگی زیاد و اندازه‌ی بلند ژنوم‌ها بدون استفاده از روش‌های محاسباتی امکان‌پذیر نیست. به طور سنتی، مونتاژ ژنوم بر پایه روش‌هایی مانند Sanger sequencing بنا نهاده شده است. این روش شامل توالی‌یابی قطعات نسبتاً طولانی DNA و استفاده از تکنیک‌هایی مثل restriction mapping و کلون‌سازی در وکتورها می‌باشد، که سپس به طور جداگانه توالی‌یابی می‌شوند. این فرآیند نیاز به یک نقشه فیزیکی دارد تا این توالی‌ها را کنار هم قرار دهد و این روند زمان‌بر و خسته‌کننده است. برای ژنوم‌های بزرگ، استفاده از روش‌های محاسباتی برای مونتاژ reads کوتاه به یک توالی ژنومی کامل ضروری است.

۱-۳ چالش از دیدگاه محاسباتی

از نظر محاسباتی، مشکل با استفاده از الگوریتم‌ها برای چیدن این readهای کوتاه به توالی‌های بلندتر (contigs) حل می‌شود. این مقاله یک تکنیک محاسباتی بهبود یافته را با ساخت گراف‌های A-Bruijn معرفی می‌کند که به جای استفاده از اطلاعات جفت readها، که با فاصله مشخصی از هم هستند، به عنوان یک گام پس‌پردازش، از این اطلاعات مستقیماً در فرآیند مونتاژ استفاده می‌شود. این روش در هدف بهبود مونتاژ نواحی ژنومی پیچیده است.

۱-۴ دیدگاه‌های حل مسئله

با توجه به پیشرفت‌های انجام‌شده در next-generation sequencing، پروژه‌هایی نظیر توالی‌یابی گونه‌های مختلف امکان‌پذیر شده است، اما چالش‌های محاسباتی به وجود آمده توسط readهای کوتاه در فرآیند مونتاژ ژنوم همچنان پابرجاست. رویکردOverlap-Layout-Consensus که با پیدا کردن همپوشانی‌ها بین readها کار می‌کند، اولین قدم‌ها را برای تکنیک‌های مونتاژ فراهم کرد. روش‌های مبتنی بر de Bruijn graphs نیز با ساختن نقشه‌هایی که از readها عبور می‌کنند، توانستند بهبودهایی در تصحیح خطاها و حل مشکل تکرارها ایجاد کنند. با این حال، مشکل تکرارهای طولانی همچنان وجود دارد که بخشی از آن با استفاده از جفت‌های mate که فواصل را پر می‌کنند، کاهش یافته است. اما، ادغام بی‌نقص اطلاعات جفت‌های mate برای مونتاژهای دقیق‌تر در مناطق غنی از تکرار ضروری است.

۱-۵ چالش‌های هر روش

در مونتاژ ژنوم، روش de Bruijn graph با مشکلاتی در مونتاژ دقیق مناطق با تکرارهای پیچیده مواجه است زیرا به طور طبیعی اطلاعات حاصل از mate pairs، جفت دنباله‌هایی که مناطق گسترده‌تری از DNA را شامل شده و می‌توانند نشان دهنده فاصله reads مختلف در سراسر ژنوم باشند، را در خود جای نمی‌دهد. این mate pairs برای پر کردن شکاف‌ها و حل ابهامات در این دنباله‌های تکراری حیاتی هستند. با این حال، روش‌های سنتی معمولاً از اطلاعات mate pair تنها پس از ساخت گراف اولیه استفاده می‌کنند، که این امر اثربخشی آن‌ها را در حل کردن این مناطق پیچیده محدود می‌کند. چالش این است که داده‌های mate pair را در طی فاز ساخت خود نمودار گنجانده شود تا دقت و پیوستگی مونتاژ بهبود یابد.

۱-۶ چالشی که توسط مقاله حل شده است

این مقاله به چالش ادغام موثر داده‌های mate pair در فرآیند مونتاژ ژنوم می‌پردازد. روش معرفی شده که این داده‌ها را هنگام ساخت گراف‌های de Bruijn ادغام می‌کند، با هدف بهبود دقت و پیوستگی ژنوم‌های مونتاژ شده، به ویژه در نواحی با تکرارهای پیچیده که مونتاژ آن‌ها با روش‌های سنتی دشوار است.

فصل دوم

الگوریتم

۲-۱ تعاریف کلیدی

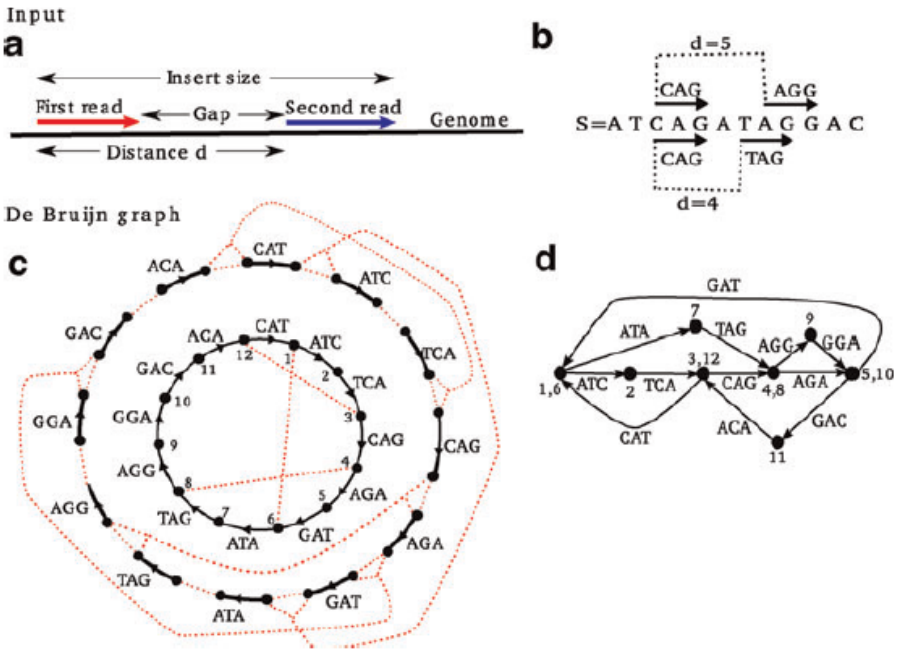
در اینجا خلاصه‌ای از تعاریف کلیدی آورده شده است:

- **Circular Genome Assumption**: فرض ساده‌سازی ژنوم به یک رشته دایره‌ای برای مدل‌سازی.
- **Uniform Read Length**: فرض می‌کند که همه خوانش‌ها دارای طول یکسان l هستند.
- **Error-Free Reads**: فرض می‌کند که خوانش‌های توالی‌یابی بدون خطا هستند.
- **Mate Pair**: یک جفت مرتب از رشته‌ها به طول l است که از ژنوم در موقعیت‌های i و j کشیده شده‌اند. و فاصله بین آنها به صورت زیر تعریف میشود.

$$d = j - i$$

- **A-Bruijn Graphs**: تغییری از گراف‌های de Bruijn که در آن رئوس بر اساس یک ماتریس از دستورالعمل‌های چسبندگی به هم متصل می‌شوند.
- **k-mer**: یک زیررشته به طول k از یک رشته دایره‌ای S است که به صورت $S(i)$ نشان داده می‌شود، که دنباله‌ای از موقعیت i تا $i+k-1$ است، با در نظر گرفتن اینکه رشته به صورت دایره‌ای است، بنابراین شاخص بر اساس n (طول S) مدولو گرفته می‌شود.
- **k-spectrum**: مجموعه‌ای از تمام k-merهای ممکن از یک رشته S، برای $1 \leq i \leq n$.
- **prefix(a)** و **suffix(a)**: برای یک k-mer به نام a، پیشوند با حذف کاراکتر آخر ($a_1 ... a_{k-1}$) و پسوند با حذف کاراکتر اول ($a_2 ... a_k$) به دست می‌آید.
- **k-mer alignment**: یک k-mer به نام a در موقعیت i در S مطابقت دارد اگر $a = S(i)$.
- **(k, d)-mers and bilabels**: یک bilabel مجموعه‌ای از دو k-mer، a و b است، که به صورت $(a|b)$ نشان داده می‌شود، که دقیقا d نوکلئوتید از هم فاصله دارند. برای یک رشته S و پارامترهای k و d، یک (k,d)-mer از S یک bilabel است که جایی در S مطابقت دارد.
- **left(a|b)** و **right(a|b)**: با توجه به یک bilabel به نام $(a|b)$، قسمت چپ a و قسمت راست b است.
- **k-mer bilabel alignment**: یک k-mer bilabel به نام $(a|b)$ در موقعیت i مطابقت دارد اگر $a = S(i)$ و $b = S(i+d+x)$ برای برخی $-\Delta \leq x \leq \Delta$، که Δ انحراف مجاز از فاصله دقیق d است.

- **Gluing Vertices**: فرایند ادغام رئوس در یک نمودار بر اساس توالی‌های همپوشانی k-mers.
- **Covering Cycle**: مسیری در نمودار که حداقل یکبار از هر لبه عبور می‌کند و کل ژنوم را به صورت توالی از همپوشانی k-mers نشان می‌دهد.



شکل ۲-۱: گراف و تعاریف فاصله دو خوانش

۲-۲ ساخت گراف de Bruijn (مدل سازی خوانش‌های جفت نشده)

شرح ساخت گراف به صورت زیر می‌باشد:

- شروع با مجموعه C از (k + 1)-mer ها از یک رشته دایره‌ای S.
- ساخت یک گراف اولیه G_0 با ایجاد رأس‌ها و لبه‌ها از (k + 1)-mer ها در C.
- برچسب‌گذاری هر لبه با (k + 1)-mer متناظر آن.
- معرفی یک ماتریس دودویی A برای نمایش دستورالعمل‌های چسباندن رأس‌ها.
- چسباندن رأس‌ها در G_0 بر اساس ماتریس A، جایی که $A_{ij} = 1$ نشان‌دهنده چسباندن رأس‌های i و j است اگر برچسب یکسانی داشته باشند.
- گراف de Bruijn نهایی $DB(C, k)$ با انجام تمام چسباندن‌های مشخص شده حاصل می‌شود، که نتیجه یک گراف ساده یا یک مولتی گراف بر اساس کثرت (k + 1)-mer ها ممکن است باشد.
- تعریف walk در گراف توسط توالی لبه‌ها، که رشته‌ها را با همپوشانی (k + 1)-mer ها با k کاراکتر نمایش می‌دهد.
- اطمینان حاصل کنید که یک covering cycle در گراف وجود دارد، که هر لبه را دست‌کم یک بار بازدید می‌کند و رشته اصلی را بازسازی می‌کند S را می‌دهد.

۲-۳ ساخت گراف Paired de Bruijn (مدل سازی خوانش‌های جفت شده با فاصله دقیق)

شرح ساخت گراف به صورت زیر می‌باشد:

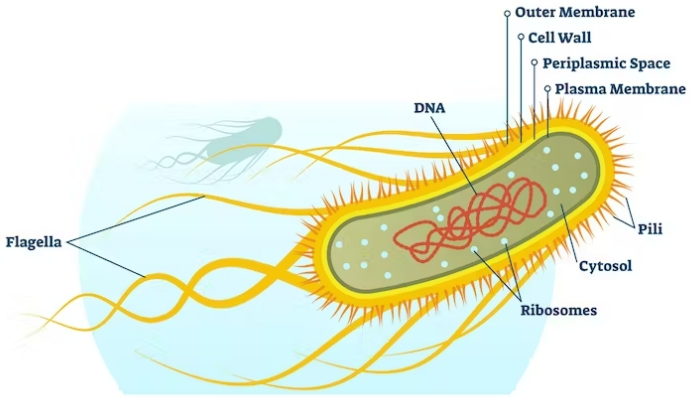
- شروع با مجموعه‌ای از (k + 1, d)-merها C از یک رشته S.
- ساختن یک گراف اولیه G_0 با ۲$|C|$ رأس، که $|C|$ تعداد (k + 1, d)-merها است.
- برای هر bilabel (a|b) در C، دو رأس جدید u و v و یک لبه $u \to v$ ایجاد کنید.
- لبه با (a|b) برچسب‌گذاری می‌شود.
- رأس u با prefix(a|b) برچسب‌گذاری می‌شود.
- رأس v با suffix(a|b) برچسب‌گذاری می‌شود.
- رأس‌ها در G_0 که برچسب یکسانی دارند را به هم بچسبانید. گراف نتیجه G، گراف de Bruijn جفتی C است.
- در گراف G، هر رأس برچسب منحصر به فردی حفظ می‌کند که مشترک بین تمام رأس‌هایی است که برای تشکیل آن به هم چسبیده‌اند.
- پیاده‌روی‌ها در G را با خاصیتی تعریف کنید که برچسب‌های چپ پیاده‌روی یک قسمت از رشته و برچسب‌های راست قسمت دیگری را با حفظ فاصله d نمایش می‌دهد.
- گراف ساخته شده G خاصیت کلیدی گراف de Bruijn را حفظ می‌کند که در آن یک چرخه پوششی وجود دارد که رشته اصلی S را نمایش می‌دهد. این چرخه برای بازسازی S و در نتیجه برای نمایش contig ضروری است.

فصل سوم

پایگاه داده

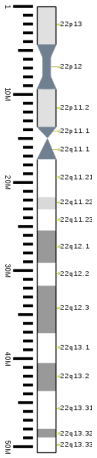
برای برسی درستی الگوریتم و رویکرد Paired de Bruijn از داده‌های E. coil[۳] و chromosome 22[۴] استفاده شده است.

E. coil مخفف Escherichia coli است که نوعی باکتری است که معمولاً در روده انسان و حیوانات یافت می شود. این یک میکروارگانیسم به خوبی مطالعه شده است و اغلب در تحقیقات علمی استفاده می‌شود.



شکل ۳-۱: E. coil

کروموزوم ۲۲ انسان یکی از ۲۳ جفت کروموزوم موجود در سلول های انسان است. این دومین کروموزوم کوچک انسان است که حدود ۵۱ میلیون جفت باز DNA را در بر می گیرد و بین ۵.۱ تا ۲ درصد از کل DNA در سلول های انسانی را تشکیل می دهد. کروموزوم ۲۲ به طور گسترده مورد مطالعه قرار گرفته است و توالی آن به عنوان بخشی از پروژه ژنوم انسانی تعیین شده است. در عملکردهای ژنتیکی مختلف نقش دارد و با اختلالات ژنتیکی خاصی همراه بوده است.



شکل ۳-۲: کروموزوم ۲۲ انسان

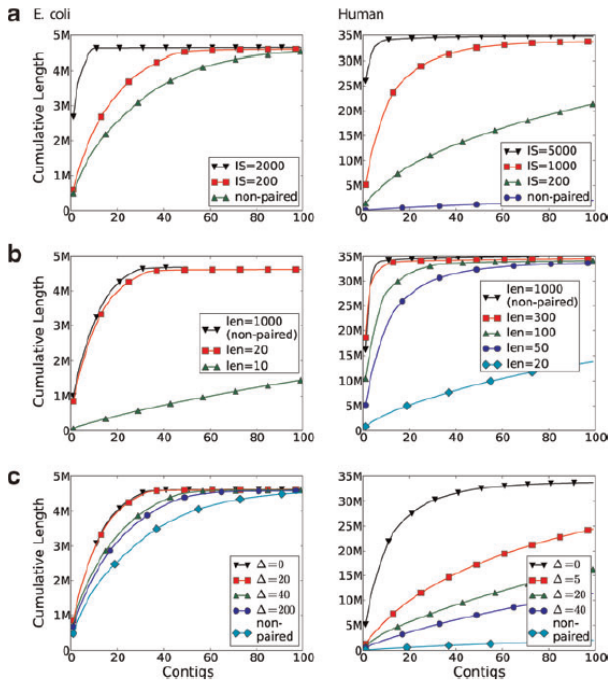
فصل چهارم

صحت الگوریتم

الگوریتم با استفاده از توالی های E. coli (4.6 Mbp) و کروموزوم انسانی 22 (35 Mbp) ارزیابی شده است و خوانش‌های شبیه سازی شده با پوشش کامل و شرایط بدون خطا تولید شده‌اند وعملکرد بر اساس توانایی الگوریتم برای تولید contigهایی اندازه گیری شده که کاملاً به ژنوم اصلی نگاشت می شدند، که این کار را با موفقیت برای همه گروه ها انجام شده است.

اثربخشی رویکرد گراف جفتی دی بروین در شرایط ایده‌آل مورد آزمایش قرار گرفت تا پتانسیل آن در بهبود اندازه‌های contig در مونتاژ ژنوم نشان داده شود. یافته‌های آن‌ها بهبود قابل‌توجهی را در طول پیوستگی با افزایش اندازه درج نشان داده شده است. به عنوان مثال، با اندازه درج 6000 nt، تمام E. coli در یک contig مونتاژ شده‌اند، و برای کروموزوم انسانی 22، اندازه درج 5000 nt امکان پوشش 98% با 15 بزرگترین contig را فراهم کرده است.

علاوه بر این، تأثیر طول‌های خوانش‌های مختلف و تغییر اندازه درج را بر کیفیت مونتاژ بررسی شده است و می‌توان دریافت دریافتند که زمانی که طول خواندن از یک آستانه کوچک فراتر می‌رود، طول‌های contig تقریباً به حد مطلوب نظری خود می‌رسند. با این حال، کیفیت مجموعه با افزایش تنوع در اندازه درج، به ویژه برای ژنوم انسان، بدتر شده است، که نشان دهنده اهمیت ثبات اندازه درج در مجموعه ژنوم است.



شکل ۴-۱: نمودارهای ارزیابی الگوریتم

- نمودارها طول های تجمعی حاصل از الگوریتم مونتاژ اعمال شده روی ژنوم E. coli و ژنوم انسان را با تمرکز بر کروموزوم 22 نشان می دهند.

- نمودار A تأثیر اندازه های مختلف درج (IS) را بر کیفیت مجموعه نشان می دهد، با طول خوانده شده در 50 ثابت شده است. یک کانتیگ منفرد که کل ژنوم E. coli را نشان می دهد با اندازه درج 6000 nt به دست می آید.

- نمودار B بررسی می‌کند که چگونه طول‌های خواندن مختلف بر طول contig تجمعی تأثیر می‌گذارند، با اندازه درج ثابت در 1000 nt. طول خواندن طولانی تر به طور قابل توجهی طول های contig را بهبود می بخشد.

- نمودار C به تأثیر تغییر اندازه درج (Δ) بر روی مجموعه نگاه می کند و میانگین اندازه درج و طول خواندن را ثابت نگه می دارد. افزایش تنوع در اندازه درج به طور کلی طول تجمعی را کاهش می دهد، که اهمیت اندازه های درج ثابت برای کیفیت مونتاژ را نشان می دهد.

- به طور کلی، این نمودارها نشان می‌دهند که هم اندازه‌های درج بزرگ‌تر و هم طول خواندن طولانی‌تر، نتایج مونتاژ را بهبود می‌بخشند، در حالی که تغییر در اندازه درج می‌تواند تأثیر مضری داشته باشد.

منابع و مراجع

[1] Commins, Jennifer, Toft, Christina, and Fares, Mario. Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. Biological procedures online, 11:52–78, 04 2009.

[2] Medvedev, Paul, Pham, Son, Chaisson, Mark, Tesler, Glenn, and Pevzner, Pavel. Paired de bruijn graphs: A novel approach for incorporating mate pair information into genome assemblers. Journal of Computational Biology, 18(11):1625–1634, 2011.

[3] Kegg genome: Escherichia coli k-12 mg1655.

[4] Ensembl - database commons.