دانشگاه صنعتی شریف

دانشکده‌ی مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی

گرایش مهندسی نرم‌افزار

عنوان:

هم‌ردیفی خوانده‌های بایسولفیت توسط آریانا

نگارش:

افسون افضل، مریم ربیعی هاشمی

استاد راهنما:

دکتر حیدرنوری

دکتر شریفی زارچی

بهمن ۱۳۹۳


بسم الله الرحمن الرحیم

به نام خدا

دانشگاه صنعتی شریف

دانشکده‌ی مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی

عنوان: هم‌ردیفی خوانده‌های بایسولفیت توسط آریانا

نگارش: افسون افضل، مریم ربیعی هاشمی

چکیده

نگارش پایان‌نامه علاوه بر بخش پژوهش و آماده‌سازی محتوا، مستلزم رعایت نکات فنی و نگارشی دقیقی است که در تهیه‌ی یک پایان‌نامه‌ی موفق بسیار کلیدی و مؤثر است. از آن جایی که بسیاری از نکات فنی مانند قالب کلی صفحات، شکل و اندازه‌ی قلم، صفحات عنوان و غیره در تهیه‌ی پایان‌نامه‌ها یکسان است، با استفاده از نرم‌افزار حروف‌چینی زی‌تک و افزونه‌ی زی‌پرشین یک قالب استاندارد برای تهیه‌ی پایان‌نامه‌ها ارائه گردیده است. این قالب می‌تواند برای تهیه‌ی پایان‌نامه‌های کارشناسی و کارشناسی ارشد و نیز رساله‌ی دکتری مورد استفاده قرار گیرد. این نوشتار به طور مختصر نحوه‌ی استفاده از این قالب را نشان می‌دهد.

کلیدواژه‌ها: پایان‌نامه، حروف‌چینی، قالب، زی‌پرشین

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

فصل ۱

نحوه‌ی نگارش

در این فصل نکات کلی در مورد نگارش پایان‌نامه به اختصار توضیح داده می‌شود.

۱-۱ پرونده‌ها

پرونده‌ی اصلی پایان‌نامه‌ی شما `thesis.tex` نام دارد. به ازای هر فصل از پایان‌نامه، یک پرونده در شاخه‌ی `chapters` ایجاد نموده و نام آن را در پرونده‌ی `thesis.tex` (در قسمت فصل‌ها) درج نمایید. پیش از شروع به نگارش پایان‌نامه، بهتر است پرونده‌ی `front/info.tex` را باز نموده و مشخصات پایان‌نامه را در آن تغییر دهید.

۱-۲ عبارات ریاضی

برای درج عبارات ریاضی در داخل متن از $...$ و برای درج عبارات ریاضی در یک خط مجزا از $$...$$ استفاده کنید. برای مثال $\sum_{k=۰}^{n} \binom{n}{k} = ۲^n$ در داخل متن و عبارت زیر

$$\sum_{k=۰}^{n} \binom{n}{k} = ۲^n$$

در یک خط مجزا درج شده است. همان‌طور که در بالا می‌بینید، نمایش یک عبارت یکسان در دو حالت درون‌خط و بیرون‌خط می‌تواند متفاوت باشد. دقت کنید که تمامی عبارات ریاضی، از جمله متغیرهای

تک‌حرفی مانند x و y باید در محیط ریاضی یعنی محصور درون علامت $ باشند.

۱-۳ علائم ریاضی پرکاربرد

برخی علائم ریاضی پرکاربرد در زیر فهرست شده‌اند.

- مجموعه‌های اعداد: $\mathbb{N}, \mathbb{Z}, \mathbb{Z}^+, \mathbb{Q}, \mathbb{R}, \mathbb{C}$
- مجموعه: $\{۱, ۲, ۳\}$
- دنباله: $\langle ۱, ۲, ۳ \rangle$
- سقف و کف: $\lceil x \rceil, \lfloor x \rfloor$
- اندازه و متمم: $|A|, \overline{A}$
- همنهشتی: $a \overset{n}{\equiv} ۱$ یا (پیمانه‌ی n) $a \equiv ۱$
- ضرب و تقسیم: $\times, \cdot, \div$
- سه‌نقطه بین کاما: $۱, ۲, \ldots, n$
- سه‌نقطه بین عملگر: $۱ + ۲ + \cdots + n$
- کسر و ترکیب: $\frac{n}{k}, \binom{n}{k}$
- اجتماع و اشتراک: $A \cup (B \cap C)$
- عملگرهای منطقی: $\neg p \vee (q \wedge r)$
- پیکان‌ها: $\to, \Rightarrow, \leftarrow, \Leftarrow, \leftrightarrow, \Leftrightarrow$
- عملگرهای مقایسه‌ای: $\neq, \leqslant, \nleqslant, \geqslant, \ngeqslant$
- عملگرهای مجموعه‌ای: $\in, \notin, \setminus, \subset, \subseteq, \subsetneq, \supset, \supseteq, \supsetneq$
- جمع و ضرب چندتایی: $\sum_{i=۱}^{n} a_i, \prod_{i=۱}^{n} a_i$

- اجتماع و اشتراک چندتایی: $\bigcup_{i=۱}^{n} A_i, \bigcap_{i=۱}^{n} A_i$
- برخی نمادها: $\infty, \emptyset, \forall, \exists, \triangle, \angle, \ell, \equiv, \therefore$

۱-۴ لیست‌ها

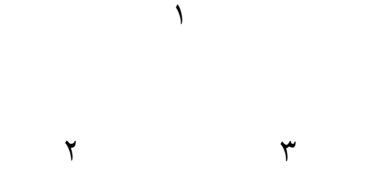
برای ایجاد یک لیست می‌توانید از محیط‌های «فقرات» و «شمارش» همانند زیر استفاده کنید.

- مورد اول
- مورد دوم
- مورد سوم

۱. مورد اول
۲. مورد دوم
۳. مورد سوم

۱-۵ درج شکل

یکی از روش‌های مناسب برای ایجاد شکل استفاده از نرم‌افزار LaTeX Draw و سپس درج خروجی آن به صورت یک فایل `tex` درون متن با استفاده از دستور `fig` یا `centerfig` است. شکل ؟؟ نمونه‌ای از اشکال ایجادشده با این ابزار را نشان می‌دهد.



شکل ۱-۱: یک گراف و پوشش رأسی آن

همچنین می‌توانید با استفاده از نرم‌افزار Ipe شکل‌های خود را مستقیما به صورت pdf ایجاد نموده و آن‌ها را با دستورات `img` یا `centerimg` درون متن درج کنید. برای نمونه، شکل ؟؟ را ببینید.

عملگر	عملیات
<	کوچک‌تر
>	بزرگ‌تر
==	مساوی
<>	نامساوی

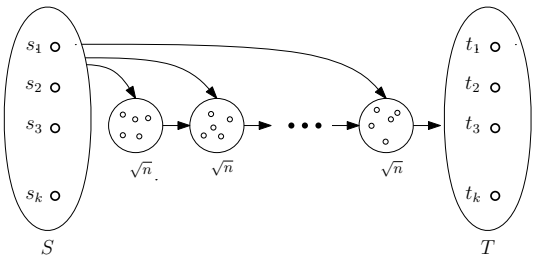
جدول ۱-۱: عملگرهای مقایسه‌ای

۱-۶ درج جدول

برای درج جدول می‌توانید با استفاده از دستور «جدول» جدول را ایجاد کرده و سپس با دستور «لوح» آن را درون متن درج کنید. برای نمونه جدول ؟؟ را ببینید.

۱-۷ درج الگوریتم

برای درج الگوریتم می‌توانید از محیط «الگوریتم» همانند زیر استفاده کنید.



شکل ۱-۲: یک گراف جهت‌دار بدون دور

الگوریتم ۱ پوشش رأسی حریصانه

ورودی: گراف $G = (V, E)$

خروجی: یک پوشش رأسی از G

۱: قرار بده $C = \emptyset$

۲: تا وقتی E تهی نیست:

۳: یال دل‌خواه $uv \in E$ را انتخاب کن

۴: رأس‌های u و v را به C اضافه کن

۵: تمام یال‌های واقع بر u یا v را از E حذف کن

۶: C را برگردان

۱-۸ محیط‌های ویژه

برای درج مثال‌ها، قضیه‌ها، لم‌ها و نتیجه‌ها به ترتیب از محیط‌های «مثال»، «قضیه»، «لم» و «نتیجه» استفاده کنید. برای درج اثبات قضیه‌ها و لم‌ها از محیط «اثبات» استفاده کنید.

تعریف‌های داخل متن را با استفاده از دستور «مهم» به صورت تیره نشان دهید. تعریف‌های پایه‌ای‌تر را درون محیط «تعریف» قرار دهید.

تعریف ۱-۱ (اصل لانه‌کبوتری) اگر $n+1$ یا بیش‌تر کبوتر درون n لانه قرار گیرند، آن‌گاه لانه‌ای وجود دارد که شامل حداقل دو کبوتر است.

فصل ۲

برخی نکات نگارشی

این فصل حاوی برخی نکات ابتدایی ولی بسیار مهم در نگارش متون فارسی است. نکات گردآوری‌شده در این فصل به هیچ وجه کامل نیست، ولی دربردارنده‌ی حداقل مواردی است که رعایت آن‌ها در نگارش پایان‌نامه ضروری به نظر می‌رسد.

۲-۱ فاصله‌گذاری

۱. علائم سجاوندی مانند نقطه، ویرگول، دونقطه، نقطه‌ویرگول، علامت سؤال، و علامت تعجب (. ، : ؛ ؟ !) بدون فاصله از کلمه‌ی پیشین خود نوشته می‌شوند، ولی بعد از آن‌ها باید یک فاصله قرار گیرد. مانند: من، تو، او.

۲. علامت‌های پرانتز، آکولاد، کروشه، نقل قول و نظایر آن‌ها بدون فاصله با عبارات داخل خود نوشته می‌شوند، ولی با عبارات اطراف خود یک فاصله دارند. مانند: (این عبارت) یا آن عبارت.

۳. دو کلمه‌ی متوالی در یک جمله همواره با یک فاصله از هم جدا می‌شوند، ولی اجزای یک کلمه‌ی مرکب باید با نیم‌فاصله[۱] از هم جدا شوند. مانند: کلاسِ درس، محبت‌آمیز، دوبخشی.

[۱] «نیم‌فاصله» فاصله‌ای مجازی است که در عین جدا کردن اجزای یک کلمه‌ی مرکب از یک‌دیگر، آن‌ها را نزدیک به هم نگه می‌دارد. معمولاً برای تولید این نوع فاصله در صفحه‌کلیدهای استاندارد از ترکیب Shift+Space استفاده می‌شود.

۲-۲ شکل حروف

۱. در متون فارسی به جای حروف «ك» و «ي» عربی باید از حروف «ک» و «ی» فارسی استفاده شود. همچنین به جای اعداد عربی مانند ٥ و ٦ باید از اعداد فارسی مانند ۵ و ۶ استفاده نمود. برای این کار، توصیه می‌شود صفحه‌کلید فارسی استاندارد[۲] را بر روی سیستم خود نصب کنید.

۲. عبارات نقل‌قول‌شده یا مؤکد باید درون علامت نقل قولِ «» قرار گیرند، نه "". مانند: «کشور ایران».

۳. کسره‌ی اضافه‌ی بعد از «ه» غیرملفوظ به صورت «ه‌ی» نوشته می‌شود، نه «هٔ». مانند: خانه‌ی علی، دنباله‌ی فیبوناچی.

تبصره: اگر «ه» ملفوظ باشد، نیاز به «ی» ندارد. مانند: فرمانده دلیر، پادشه خوبان.

۴. پایه‌های همزه در کلمات، همیشه «ئـ» است، مانند: مسئله و مسئول، مگر در مواردی که همزه ساکن است که در این صورت باید متناسب با اعراب حرف پیش از خود نوشته شود. مانند: رأس، مؤمن.

۲-۳ جدانویسی

۱. اجزای فعل‌های مرکب با فاصله از یک‌دیگر نوشته می‌شوند، مانند: تحریر کردن، به سر آمدن.

۲. علامت استمرار، «می»، توسط نیم‌فاصله از جزء بعدی فعل جدا می‌شود. مانند: می‌رود، می‌توانیم.

۳. شناسه‌های «ام»، «ای»، «ایم»، «اید» و «اند» توسط نیم‌فاصله، و شناسه‌ی «است» توسط فاصله از کلمه‌ی پیش از خود جدا می‌شوند. مانند: گفته‌ام، گفته‌ای، گفته است.

۴. علامت جمع «ها» توسط نیم‌فاصله از کلمه‌ی پیش از خود جدا می‌شود. مانند: این‌ها، کتاب‌ها.

۵. «به» همیشه جدا از کلمه‌ی بعد از خود نوشته می‌شود، مانند: به نام و به آن‌ها، مگر در مواردی که «بـ» صفت یا فعل ساخته است. مانند: بسزا، ببینم.

[۲] صفحه‌کلید فارسی استاندارد برای ویندوز، تهیه‌شده توسط بهنام اسفهبد

۶. «به» همواره با فاصله از کلمه‌ی بعد از خود نوشته می‌شود، مگر در مواردی که «به» جزئی از یک اسم یا صفت مرکب است. مانند: تناظر یک‌به‌یک، سفر به تاریخ.

۲-۴ جدانویسی مرجح

۱. اجزای اسم‌ها، صفت‌ها، و قیدهای مرکب توسط نیم‌فاصله از یک‌دیگر جدا می‌شوند. مانند: دانش‌جو، کتاب‌خانه، گفت‌وگو، آن‌گاه، دل‌پذیر.

تبصره: اجزای منتهی به «هاء ملفوظ» را می‌توان از این قانون مستثنی کرد. مانند: راهنما، رهبر.

۲. علامت صفت برتری، «تر»، و علامت صفت برترین، «ترین»، توسط نیم‌فاصله از کلمه‌ی پیش از خود جدا می‌شوند. مانند: بیش‌تر، کم‌ترین.

تبصره: کلمات «بهتر» و «بهترین» را می‌توان از این قاعده مستثنی نمود.

۳. پیشوندها و پسوندهای جامد، چسبیده به کلمه‌ی پیش یا پس از خود نوشته می‌شوند. مانند: همسر، دانشکده، دانشگاه.

تبصره: در مواردی که خواندن کلمه دچار اشکال می‌شود، می‌توان پسوند یا پیشوند را جدا کرد. مانند: هم‌میهن، هم‌ارزی.

۴. ضمیرهای متصل چسبیده به کلمه‌ی پیش از خود نوشته می‌شوند. مانند: کتابم، نامت، کلامشان.

فصل ۳

مقدمه

متیلاسیون[۱] سیتوزین[۲] از بسیاری از جهات بر بیولوژی انسان تأثیرگذار است از جمله رشد جنینی، رونویسی[۳]، ساختار کروماتین[۴] و ... . این مورد در گیاهان نیز به همان اندازه، در مواردی چون رونویسی، ترمیم DNA[۵] و تفاوت سلولی[۶]، اهمیت دارد. نکته قابل توجه آن است که متیلاسیون DNA در انعطاف و حافظه سیستم عصبی مؤثر است و همچنین متیلاسیون غیر طبیعی DNA عامل بسیاری از بیماری‌ها از جمله آلزایمر و سرطان است. روش‌های درمانی سرطان در حال توسعه هستند که با هدف یافتن الگوی نرمال متیلاسیون عمل می‌کنند. روش استاندارد اندازه‌گیری متیلاسیون DNA، ترکیب نمونهٔ برداشت شده با سدیم بایسولفیت است که سیتوزین‌های غیرمتیله را به یوراسیل (که پس از تکثیر به تیامین تبدیل می‌شود) تبدیل می‌کند. پس از آن DNA، توالی‌یابی می‌گردد و با ژنوم مرجع مقایسه می‌شود به طوری که نگاشت C به C نشان‌دهندهٔ متیله بودن و نگاشت T به C نشان‌دهندهٔ غیر متیله بودن است. روش‌ها و الگوریتم‌های گوناگونی برای توالی‌یابی خوانده‌های بایسولفیت‌شده ارائه شده‌اند و بر اساس آنها ابزارهایی توسعه یافته‌اند. با این حال ضعف این ابزارها در کمی دقت و همچنین عدم درنظرگیری خواص زیستی است. به همین دلیل ما تلاش نمودیم که با توسعهٔ ابزار آریانا، که یک

[۱] Methylation

[۲] Cytosine

[۳] transcription

[۴] chromatin structure

[۵] DNA repair

[۶] DNA repair

توالی‌یاب قدرتمند نسل جدید است، و در نظرگیری خواص زیستی از جمله خواص CpG ها و جزایر CpG دقت را افزایش دهیم.

۳-۱ تعریف مسئله

در این پروژه ما سعی داشتیم که یک aligner برای خوانده‌های treated bisulfite بنویسیم . هدف مساله، نگاشت یک تعداد از رشته‌های متشکل از حروف A،T،C،G بر روی یک ژنوم خاص می‌باشد که تمامی این رشته‌ها از یک ژنوم بدست آمده‌اند ولی کاملا مشابه آن نیستند و تغییراتی در آنها صورت پذیرفته است. هدف یافتن جایگاه این رشته‌ها در ژنوم با دقت حداکثر و درنظر داشتن محدودیتهای زمانی و حافظه‌ای با توجه به ابعاد مساله است . در این پروژه سعی ما بر آن بود که با گسترش ابزار آریانا ، قابلیت هم‌ردیفی خوانده‌های treated bisulfite با ژنوم مرجع را به آن بیفزاییم . همانطور که توضیح داده شد در داده‌های بایسولفیت ، سیتوزین‌هایی که متیله نباشند به تیامین تبدیل شده‌اند و بنابراین با ژنوم مرجع متفاوتند . به دلیل اهمیت نقش متیلاسیون و نبود یک راه حل دقیق، تلاش‌های بسیاری برای حل این مسئله شده است(Kruger،۲۰۱۲). سعی ما در این پروژه بر آن بود که راه‌حلی که ارائه می‌کنیم بر خلاف سایر aligner ها خواص زیستی موثر در متیلاسیون را در نظر بگیرد و با استفاده از این ویژگی بتواند دقت هم‌ردیفی را افزایش دهد.

۳-۲ اهمیت موضوع

مسئله‌ی مسیریابی وسایل نقلیه کاربردهای بسیار گسترده‌ای در حوزه‌ی حمل و نقل دارد. برای نخستین بار این مسئله برای مسیریابی تانکرهای سوخت‌رسان مطرح شد [؟]. اما امروزه با پیشرفت‌های گسترده‌ای که در زمینه‌ی تکنولوژی روی داده است از راه‌حل‌های این مسئله در امور روزمره از جمله سیستم توزیع محصولات، تحویل نامه، جمع‌آوری زباله‌های خانگی و غیره استفاده می‌شود. در نظر گرفتن فرض ناهمگن بودن هم با توجه به اینکه معمولاً عوامل توزیع در یک سیستم، یکسان نیستند و تفاوت‌هایی در میزان مصرف سوخت و غیره دارند، راه‌حل‌های مناسب‌تری برای مسائل این حوزه می‌تواند ارائه دهد. گونه‌های مختلفی از مسائل مسیریابی وسایل نقلیه در [؟، ؟، ؟] بیان شده است.

۳-۳ اهداف تحقیق

در این پایان‌نامه سعی می‌شود که مسئله‌ی توالی‌یابی خوانده‌های بایسولفیت‌شده, به کمک ابزار آریانا و با توجه به خواص زیستی اثبات‌شده مورد بررسی قرار گیرد و راه حلی کارا برای آن ارائه شود.

۳-۴ ساختار پایان‌نامه

این پایان‌نامه شامل پنج فصل است. فصل دوم دربرگیرنده‌ی تعاریف اولیه‌ی مرتبط با پایان‌نامه است. در فصل سوم مسئله‌ی دورهای ناهمگن و کارهای مرتبطی که در این زمینه انجام شده به تفصیل بیان می‌گردد. در فصل چهارم نتایج جدیدی که در این پایان‌نامه به دست آمده ارائه می‌گردد. در این فصل، مسئله‌ی درخت‌های ناهمگن در چهار شکل مختلف مورد بررسی قرار می‌گیرد. سپس نگاهی کوتاه به مسئله‌ی مسیرهای ناهمگن خواهیم داشت. در انتها با تغییر تابع هدف، به حل مسئله‌ی کمینه کردن حداکثر اندازه‌ی درخت‌ها می‌پردازیم. فصل پنجم به نتیجه‌گیری و پیش‌نهادهایی برای کارهای آتی خواهد پرداخت.

فصل ۴

مفاهیم اولیه

در این فصل به تعریف مفاهیمی می‌پردازیم که در پایان‌نامه مورد استفاده قرار گرفته‌اند.

۴-۱ رشته‌ی DNA

دی ان ای مولکولی است که اطلاعات ژنتیکی مورد نیاز برای رشد و فعالیت همه ارگانیسم‌ها و برخی از ویروسها را encode می‌کند. دی ان ای نخستین بار در سال ۱۸۷۰ توسط فردریک میشر از هسته سلول استخراج و شناسایی گردید. دی ان ای ساختار دو رشته‌ای دارد و این دو رشته مانند زیپ به هم متصل شده و حول یک محور مشترک پیچیده شده‌اند. دی ان ای یک پلیمر بسیار طویل است که از تکرار واحد هایی به نام نوکئوتید به دست می‌آید. هر نوکلئوتید از یک باز نوکلئوتیدی شامل نیتروژن تشکیل شده است که این بازها آدنین (A) ، تیامین (T) ، سیتوزین (C) ویا گوانین (G) که در زیر شکل ساختار آنها مشاهده می‌شود:

نوکلئوتیدها در یک زنجیره توسط پیوندهای کوالانسی بین قند یک نوکلئوتید و فسفات نوکئوتید بعدی به یکدیگر متصل شده‌اند. بر اساس قوانین pairing base پیوند هیدروژنی دو نوکلئوتید از دو strand رشته دی ان ای را به هم متصل میکند. این پیوند به صورتی برقرار می‌گردد که ادنین و تیامین وهمینطور سیتوزین و گوانین روبروی هم قرار بگیرند.

۲-۴ همانندسازی DNA

برای اینکه وراثت امکان‌پذیر باشد ، ژنها باید توانایی همانندسازی داشته باشند. ژن ها هر زمان که سلول تقسیم می‌شود باید کپی شوند و هر یک از دو سلول فرزند یک کپی از اطلاعات زیستی والد را دریافت می‌کنند. در همانند سازی ،DNA دو رشته آن به کمک آنزیم هلیکاز مانند زیپ از یکدیگر جدا می‌شوند و سپس از روی هر رشته، رشته‌ی جدیدی ساخته می‌شود. کلید ساخت رشته جدید در این است که روبروی هر باز ، باز مکمل آن قرار می‌گیرد، به این ترتیب با استفاده از نوکلئوتیدهای آزاد که در سیتوپلاسم وجود دارند، در مقابل ،A باز T ودرمقابل C بازG قرار می‌گیرد و در آخر دو کپی کاملا مشابه والد ساخته می‌شود.

۳-۴ ژنوم

اطلاعات ژنتیکی حمل شده توسط DNA در توالی وترتیب خطی DNA به واحدهای عملکردی مجزایی به نام ژنها تقسیم شده است که به طور شاخص حدود ۵۰۰۰تا ۱۰۰۰۰۰ نوکلئوتید طول دارند .به مجموعه این ژن ها ژنوم گفته می شود(Waterman،۱۹۹۵).

۴-۴ خوانده

مولکول دی ان ایی که در سلولهای زنده موجود است بسیار بزرگ است و ممکن نیست که چنین مولکول بزرگی در تنها یک آزمایش بدست آید. استراتژی فعلی توالی‌یابی رشته دی ان ای این است که مولکول بزرگ آن به قطعات کوچک شکسته شود و هرکدام از آنها جداگانه توالی‌یابی شوند. به این قطعات fragment یا read (خوانده)می‌گویند.

۵-۴ خوانده‌های بایسولفیت‌شده

شامل خوانده‌هایی است که در محلول سدیم بایسولفیت قرار گرفته‌اند که طی این عمل سیتوزین‌های غیر متیله، ابتدا به یوراسیل و سپس به تیامین تبدیل می‌شوند. در این بین مفاهیم دیگری مطرح می‌شوند

که به شرح زیر است:

۱. context: CpG context CpG به معنی سیتوزین‌هایی می‌باشد که بلافاصله بعد از آن G آمده است. بعد از یک سیتوزین در یک رشته، هم می‌تواند T A، و یا C بیاید. بنابراین تمام سیتوزین‌ها در CpG هاcontext نمی‌باشند ولی تمام هاCpG به طور مشخص در CpG ها هستند(Gardiner،۱۹۸۷).

۲. islands: CpG تعریف کاربردی از island CpG ها به یک ناحیه به همراه حداقل ۲۰۰ جفت باز برمی گرددکه درصد تعداد هاCG بیشتر از ٪۵۰ باشد و نرخ مشاهده شده مورد انتظار CpG بیشتر از ٪۶۰ باشد. منظور از نرخ مشاهده شده مورد انتظار عددی است که از فرمول (۱-۲) بدست می آید(Gardiner،۱۹۸۷).

File Sam ۶-۴

islands: CpG تعریف کاربردی از island CpG ها به یک ناحیه به همراه حداقل ۲۰۰ جفت باز برمی گرددکه درصد تعداد هاCG بیشتر از ٪۵۰ باشد و نرخ مشاهده شده مورد انتظار CpG بیشتر از ٪۶۰ باشد. منظور از نرخ مشاهده شده مورد انتظار عددی است که از فرمول (۱-۲) بدست می آید(Gardiner،۱۹۸۷).

رشته‌ی CIGAR ۷-۴

رشته‌ای است برای مشخص کردن اینکه کدام باز خوانده با کدام باز ژنوم مرجع همردیف شده است . این رشته از حروف i،d،m و اعداد طبیعی تشکیل شده است که ترکیب هر حرف و عدد نشان دهنده به ترتیب تعداد درج‌های متوالی ، تعداد حذف‌های متوالی و تعداد تطابق / عدم تطابق‌های متوالی است و این جفت های عدد و کاراکتر به صورت پشت سر هم ظاهر می‌شوند.

۸-۴ توالی‌یابی نسل بعد

تکنولوژی Next Generation DNA Sequencing Technology) NGS Technique) مجموعه ای از تکنیک های جدید برای خواندن توالی DNA می‌باشد که از نظر دقت افزایش و هزینه، کاهش قابل توجهی نسبت به تکنیک‌ها و متدهای قبلی دارد. در این تحقیق از داده‌های بدست آمده از این تکنولوژی، استفاده می‌شود(Ansorge، ۲۰۰۹).

۹-۴ Alignment Sequence

یک هم‌ردیفی توالی, روشی برای پشت هم قرار دادن توالی‌های DNA، RNA یا پروتئین است تا مناطق شباهت که ممکن است علت روابط رفتاری, ساختاری و یا تکاملی میان توالی‌ها باشند را شناسایی کند.

۱۰-۴ Sequencing DNA

همانطور که گفته شد، DNA یک زنجیره خطی از چهار نوکلوتید می‌باشد. اطلاعات ژنتیکی DNA در توالی این نوکلئوتیدها، رمز شده است. پروسه تعیین توالی نوکلئوتیدها در مولکول DNA راDNA sequencing گویند. متدها و تکنولوژی‌های مختلفی برای یافتن توالی DNA استفاده شده است (Waterman، ۱۹۹۵).

متیلاسیون DNA شامل اضافه شدن یک گروه متیل به انتهای کربن سیتوزین، C۵، توسطDNA methyltransferases می‌باشد. به سیتوزینی که گروه متیل به آن اضافه شده است متیل سیتوزین گفته می‌شود(شکل۲-۱۰) .(Krueger، ۲۰۱۲) http://www.ncbi.nlm.nih.gov/pubmed/۲۲۲۹۰۱۸۶

۱۱-۴ آریانا

آریانا یک برنامهٔ هم‌ردیف‌ساز است که از الگوریتم Wheeler Burrows برای نگاشت خوانده‌ها به ژنوم مرجع استفاده می‌کند. از نقاط برتری این هم‌ردیف‌ساز سرعت و دقت بالا است که در مقایسه با هم‌ردیف‌سازهای موجود قابل توجه است. به خصوص نتایج آزمایشات انجام شده نشان می‌دهد با

افزایش طول توالی‌ها سرعت آریانا نسبت به سایر روش‌های مورد بررسی برتری دارد. با توجه به اینکه با پیشرفت روش‌های توالی‌سازی طول توالی‌های ساخته شده روز به روز بلندتر می‌شوند این برتری اهمیت بیشتری پیدا می‌کند.

فصل ۵

کارهای پیشین

روش‌های بسیاری برای هم‌ردیفی خوانده‌های بایسولفیت, مانند روش‌های wild» «card و «سه حرفی» معرفی شده‌اند. دو نوع پیاده‌سازی برای روش wild» «card وجود دارد:

۱. در این روش به تمامی سیتوزین‌ها و تیامین‌های خوانده این اجازه داده می‌شود که به سیتوزین ژنوم مرجع نگاشت شوند.

۲. در روش دوم تمامی ترکیبات سیتوزین و تیامین را برای هر طول seed می‌شمارد و سپس با روش‌های hashing آن را نگاشت می‌کند.

در روش «سه حرفی» تمامی سیتوزین‌ها در خوانده‌ها و در ژنوم مرجع به تیامین تبدیل می‌شوند. در هر دوی این روش‌ها می‌توان نگاشت gapped و یا ungapped را بسته به برنامهٔ مورد استفاده, پیاده‌سازی کرد.

۵-۱ الگوریتم های ابتدایی

برنامهٔ CokusAlignment از روشی برای نگاست خوانده‌ها به ژنوم Arabidopsis که بر اساس الگوریتم‌های جستجوی درخت بنا شده است که هم از نظر محاسبات و هم از نظر حافظه بسیار ضعیف است. CokusAlignment با سرعت متوسط ۲۵ reads/sec/CPU با ژنوم تقریبا کوچک اجرا

می‌شود. لازم به ذکر است که می‌توان با بهینه‌سازی برای پروژه‌های مختلف این سرعت را بهبود بخشید اما در بسیاری از پروژه‌ها چنین کاری امکان‌پذیر نیست. همچنین این روش برای پیاده‌سازی نگاشت‌های gapped و pair-end نیز مناسب نیست. از نظر عملی, به دلیل کمبود سرعت و عملکرد, این روش قابل استفاده نیست.

۵-۲ ابزار بیسمارک

هدف ابزار ،Bismark یافتن یک تطابق منحصر به فرد با چهار بار اجرای پردازش‌ها به صورت هم‌زمان می‌باشد. در ابتدا هایread بایسولفیت با تغییراتی از نوع C به T (سیتوزین به تیامین)و از نوع G به گوانین)A به آدنین) تبدیل شده است (برابر با تغییرات سیتوزین به تیامین در رشته معکوس). سپس هر کدام از آنها به صورت معادل از نوع‌های پیش تغییر یافته از ژنوم رفرنس نگاشت می‌شوند که این عمل با استفاده از نرم افزار نگاشت Bowtie و به صورت چهار نمونه موازی صورت می گیرد. این نگاشت‌ها، Bismark را قادر می‌سازد تا به صورت مشخص، رشته اصلی بایسولفیت را مشحص نماید(Kruger،۲۰۱۱) خروجی نگاشت ابتدایی Bismark شامل یک خط برای هر read و یک تعداد از اطلاعات مفید مانند جایگاه نگاشت، رشته رفرنس که به آن نگاشت شده و read بایسولفیت شده نگاشت می‌باشد. این اطلاعات را می‌توان به عنوان پس‌پردازش درنظر گرفت. همچنین Bismark خروجی متیلاسیون را در بین ناحیه های مختلف CPG،CHG،CHH در نظر می‌گیرد و این ناحیه‌ها را از هم جدا در نظر می گیرد

۵-۳ ابزار BSmap

BSmap از این قضیه اصلی که تمام جایگاه‌های سیتوزین که تغییرات نا متقارن C/T (سیتوزین به تیامین) در آنها رخ می‌دهند ، شناخته شده هستند، برای راهنمایی در هم‌ردیفی هایread بایسولفیت شده استفاده می کند. ،BSMAP تیامین‌های موجود درخوانده‌های بایسولفیت شده را به عنوان سیتوزین mask می نماید.(برعکس تغییرات بایسولفیت)که این تغییر را فقط در جایگاه های سیتوزین ژنوم اصلی انجام می دهد درحالیکه کل تیامین‌های دیگر را در خوانده‌های بایسولفیت شده بدون تغییر نگه می دارد.بعد از انجام این تغییرات ،خوانده‌های بایسولفیت شده را به طور مستقیم بر روی ژنوم نگاشت می

نماید .(Chen،۲۰۱۰) علاوه بر موارد ذکر شده،BSMAP براساس الگوریتم کاراترtable HASH seeding کار می‌کند به این صورت که ژنوم مرجع را برای تمام هایk-mer ممکن شاخص گذاری می نماید که هر کدام از آنها seed خوانده می‌شود. برای یافتن یک نگاشت، تنها جایگاه‌هایی به(seed) طورکامل با بخشی ازخوانده منطبق شده اند جست و جو می‌شوند. با جست و جو در جدول ،هاseed اکثریت جایگاه‌های غیرنگاشت شده دورریخته می‌شوند و کارایی جست و جو به صورت قابل ملاحظه ای افزایش می یابد(Chen،۲۰۱۰).

۵-۴ کاستی‌های روش‌های پیشین

این روش خواده‌های بایسولفیت را تغییر نمی‌دهد بلکه mismatch سیتوزین به تیامین را مجاز می‌داند. بزرگ‌ترین کاستی این روش آن است که تعداد نگاشت‌های سیتوزین/تیامین که در یک خوانده می‌تواند شناسایی شود محدود به تعداد هایmismatch مجاز در نرم‌افزار مورد استفاده است. این تعداد می‌تواند توسط هایmismatch واقعی (هاSNP) بیشتر کاسته شود و نتیجه را بیش از قبل محدود کند.

فصل ۶

راه حل

۶-۱ الگوریتم

روشی که برای هم‌ردیفی خوانده‌های بایسولفیت در aligner های امروزی مانند Bismark و Brat و Brat-BW رایج است تبدیل تمامی سیتوزین‌های ژنوم مرجع و تمامی سیتوزین‌های خوانده به تیامین و پس از آن هم‌ردیف کردن خوانده‌ها با ژنوم رفرنس با استفاده از bowtie یا روشهای hashing است . مشکلی که در این روش وجود دارد این است که این aligner ها دقت کافی با توجه به اهمیتِ پیدا کردن نقاط متیله شده پیدا نکرده‌اند و در این‌ها برای بالا بردن دقت هم‌ردیفی و پیدا کردن مکان درست متناظر با هر خوانده در ژنوم مرجع ، محدودیت وجود دارد . مشکلی که در این روشها وجود دارد در این نکته نهفته است که در هم‌ردیف کردن سیتوزین با تیامین چهار حالت ممکن است رخ بدهد:

- هم‌ردیفی یک سیتوزین در ژنوم با یک سیتوزین در خوانده
- هم‌ردیفی یک سیتوزین در ژنوم با یک تیامین در خوانده
- هم‌ردیفی یک تیامین در ژنوم با یک تیامین در خوانده
- هم‌ردیفی یک تیامین در ژنوم با یک سیتوزین در خوانده

سه حالت اول برای خوانده‌های بایسولفیت امکان‌پذیر است ؛ اگر سیتوزین خوانده متیله باشد حالت اول پیش می‌آید ، اگر متیله نباشد حالت دوم و حالت سوم نیز هم‌ردیفی T با T است و مشکلی ندارد.

حالت آخر در هیچ شرایطی امکان‌پذیر نیست و این هم‌ردیفی قابل قبول نمی‌تواند باشد ولی باتوجه به اینکه در روش گفته شده تمامی سیتوزین‌ها چه در خوانده و چه در ژنوم به تیامین تبدیل می‌شوند ، حالت چهارم از بقیه حالات قابل تمییز دادن نیست و به عنوان یک هم‌ردیفی قابل قبول در این روش‌ها پذیرفته می‌شود.

تغییرات بایسولفیت و تبدیلات نامتقارن سیتوزین‌ها، فضای جست و جو را به صورت قابل توجهی افزایش می‌دهد. رشته های وcrick Watson که در واکنش با بایسولفیت تغییر یافته‌اند دیگر مکمل یکدیگر نمی باشند زیرا تغییرات بایسولفیت فقط در سیتوزین‌های غیرمتیله رخ می‌دهد.

نگاشت سیتوزین‌ها به تیامین‌ها به صورت نامتقارن انجام می‌شود. تیامین موجود در هایread بایسولفیت شده هم می تواند به سیتوزین و هم می تواند به تیامین در رفرنس ژنوم نگاشت شود اما این امر به صورت عکس امکان پذیر نمی‌باشد. این مشکل نه تنها فضای جست و جو را افزایش می‌دهد بلکه پروسه نگاشت را نیز پیچیده تر می نماید. مانند مثالی که در ادامه آورده شده است(شکل ۲-۴).رشته ATTTCG به سه رشته متفاوت قابل نگاشت است که فضای جست و جو را افزایش داده است(Xi،۲۰۰۹)

تفکر در این موضوع ما را بر این داشت که روشی جدید ارائه کنیم که در آن ناآگاهانه ژنوم و خوانده‌ها را تغییر ندهیم و عوامل زیستی ناحیه‌ای که یک باز در آن قرار گرفته و مشاهداتی که از خواص این نواحی در مورد متیله شدنشان بدست آمده است ، را در ایجاد تغییرات در ژنوم برای تطابق بیشتر با خوانده‌های بایسولفیت اعمال کنیم . سیتوزین های context CpG با احتمال ٪۹۰ متیله هستند و سیتوزین های خارج این نواحی در حدود ۱ درصد. به علت اهمیت بالای متیلاسیون این نواحی اهمیت بسیاری پیدا می‌کنند . از طرف دیگر جزایر CpG که نواحی‌ای هستند که نوکلئوتیدهایCpG آنها چگالی بالایی دارند و سیتوزینهای این هاCpG از قاعده فوق مستثنا هستند و درصد متیلاسیون در این نواحی بسیار پایین است. ما در این روش ژنوم‌های مختلفی برای پشتیبانی حالات مختلفی که می‌توان برای سیتوزین‌ها در نظر گرفت ، تولید می‌کنیم . به جای آنکه تمامی سیتوزین‌ها را به تیامین تبدیل کنیم به صورت انتخابی و با توجه به مکانی که آن سیتوزین در آن قرار گرفته است این کار را انجام می‌دهیم ؛ به این صورت که فرض را بر این می‌گذاریم که سیتوزین‌هایی که در نواحی context CpG قرار دارند متیله هستند بنابراین پس از بایسولفیت شدن تغییری نمی‌کنند و سیتوزین‌هایی که در خارج از این نواحی هستند با احتمال خوبی متیله نیستند و پس از بایسولفیت شدن به تیامین تبدیل می‌شوند. همچنین با توجه به اینکه سیتوزینهای درون جزایر CpG با احتمال بیشتری متیله نیستند ، سیتوزین‌هایی که درون نواحی context CpG و

جزایر CpG هستند را به همراه سیتوزین‌های بیرون context CpG به تیامین تبدیل کردیم و بقیه بدون تغییر باقی ماندند. ژنومی که با تغییرات فوق بدست می‌آید برای هم‌ردیفی تمامی خوانده‌های بایسولفیت کافی نیست . علت آن این است که در بدست آوردن خوانده‌ها حتی با روش‌های generation Next sequencing خطا وجود دارد و ممکن است که در خوانده‌ای تیامین به اشتباه سیتوزین گرفته شود (چه فعلی جز خوانده شود؟) و در فرآیند هم‌ردیفی آن خوانده مشکل بوجود آورد . دلیل دیگر این است که در سه حالت فوق ما فرض کردیم که تمامی سیتوزینهای خوانده‌های بیرون context CpG و یا درون جزایر CpG متیله نیستند و تمامی خوانده‌های درون ناحیه و خارج از جزایر متیله‌اند ولی در واقعیت ما این قطعیت را نداریم . برای جبران این تفاوت ما علاوه بر ژنوم فوق ژنوم دیگری تولید می‌کنیم و در آن تمامی سیتوزین‌ها را به تیامین تبدیل می‌نماییم . بدین صورت خوانده‌هایی که با ژنوم قبل نتوانند هم‌ردیف شوند با این ژنوم خواهند شد. برای اینکه هم‌ردیفی با خواص زیستی تطابق بیشتری داشته باشد در انتخاب هم‌ردیفی نهایی به ژنوم قبل اولویت داده می‌شود. علاوه بر این دو ژنوم نیاز این دیده می‌شود که برای اینکه خوانده‌های بیشتری را بتوانیم همردیف کنیم ، خوانده‌ها با ژنوم اصلی نیزهمردیف شوند . دلیل آن هم این است که ممکن است نواحی‌ای وجود داشته باشند که سیتوزینها علی‌رغم اینکه CpG نیستند کاملا متیله باشند و ظاهراین نواحی مشابه ژنوم اصلی باشد. دلیل بعدی این موضوع این است که treatment bisulfite روی همه خوانده ها کامل عمل نمی‌کند و ممکن است یک سیتوزین غیر متیله به تیامین تبدیل نشود . برای پوشش دادن این حالات نیز بهتر است که از ژنوم اصلی برای هم‌ردیفی استفاده شود. علاوه بر ژنوم‌های فوق, لازم دیدیم که ژنوم دیگری را نیز در نظر بگیریم که در آن CpG های درون جزیره را مانند دیگر CpG ها متیله در نظر بگیریم. به این دلیل که متیله نبودن CpG ها در جزیره یک رخداد قطعی نیست و این امکان وجود دارد که خوانده‌ای وجود داشته باشد که مطابق انتظار ما نباشد. تلاش ما بر این بود که تا جای ممکن تمامی حالات ممکن برای خوانده‌ها را مدنظر قرار بدهیم. لازم به ذکر است که پس از انجام آزمایش‌های فراوان با خوانده‌های شبیه‌سازی‌شده و واقعی, نتیجهٔ لازم را از این ژنوم اضافه را دریافت نکردیم و وجود آن را بی‌مورد احساس کرده و آن را از فرآیند حذف نمودیم. (عکس, نمودار, هرچی!) این نکته قابل توجه است که در aligner های موجود, مرسوم است که علاوه بر ژنوم, تمامی سیتوزین‌ها در خوانده نیز به تیامین تبدیل شده و سپس هم‌ردیفی انجام می‌شود اما در راه حلی که ما ارائه کرده‌ایم چنین عملی بی‌مورد به نظر می‌رسد. باید توجه داشت که اکثر خوانده‌هایی که هم‌ردیفی ناموفقی با ژنوم ساخته شده در قسمت قبل داشته‌اند, خوانده‌هایی هستند که سیتوزین‌های آن‌ها متیله نبوده و به تیامین تبدیل شده‌اند ولی ما در ژنوم ساخته‌شده به اشتباه آن‌ها را

سیتوزین نگه داشته‌ایم. در این حالت دیگر نیازی به تغییر در خوانده نیست و باید بدون تغییر به ژنومی که تمامی سیتوزین‌های آن به تیامین تبدیل شده است, هم‌ردیف گردد.

******** یکی از موارد بسیار مهم در زمینهٔ هم‌ردیفی خوانده‌های بایسولفیت‌شده, توان‌مندی aligner در هم‌ردیفی خوانده‌های PCR شده است. می‌توان نشان داد که در این حالت, ژنوم‌های مطرح شده کافی نیستند و خوانده‌های مورد نظر را پوشش نمی‌دهند. همانطور که در شکل زیر پیداست, در این نوع, خوانده‌ها هم از رشتهٔ مثبت و هم از رشتهٔ منفی جمع‌آوری می‌گردند و در دستگاه PCR تکثیر می‌شوند. سیتوزین‌های غیر متیله در این خوانده‌ها, پس از بایسولفیته شدن به تیامین تبدیل می‌شوند. به همین دلیل, اگر خوانده‌ای از رشتهٔ منفی برداشته شده باشد, پس از بایسولفیته شدن معادل ان خواهد بود که در معکوس این خوانده گوانین‌های غیرمتیله به آدنین تبدیل شده باشند. بدیهی‌ست که این مورد را نمی‌توان با ژنوم‌های پیشتر مطرح‌شده پوشش داد. به همین دلیل ژنوم‌هایی ساخته می‌شود که در آنها در رشتهٔ مثبت, گوانین‌ها به آدنین تبدیل می‌گردد. البته ما همچنان خواص زیستی را در این مورد نیز مورد توجه قرار داده‌ایم و برای ساخت این ژنوم مناطق CG و مناطق جزیرهٔ CpG را مد نظر قرار می‌دهیم. (عکس, شکل)

۶-۲ انتخاب بهترین هم‌ردیفی

به منظور انتخاب بهترین هم‌ردیفی برای هر خوانده از میان هم‌ردیفی‌هایی مختلفی که می‌تواند با ژنومهای تولید شده داشته باشد ، به محاسبه پنالتی می‌پردازیم . به این صورت که باز به باز خوانده و ژنوم مرجع را در نظر گرفته و بر اساس جایگاه باز و نوع عدم تطابق، یک پنالتی از میان سه پنالتی کم ، متوسط و زیاد که مقادیر آنها ورودی برنامه هستند در نظر می‌گیریم . پنالتی یک خوانده در حال حاضر جمع پنالتی‌های بازهای آن می‌باشد ولی برنامه نوشته شده به این صورت است که قابلیت تغییر تابع محاسبه پنالتی به سادگی وجود دارد و می‌توان مدل آماری مناسب برای این کار را بدست آورد و با کمترین تغییرات ممکن آن را به کد اضافه کرد. به دلیل اینکه هر خوانده با تمامی ژنومها هم‌ردیف می‌شود ، این نیاز احساس می‌شود که برای مقادیر پنالتی که برای خوانده‌ها بدست می‌آید سقف در نظر بگیریم تا به بهای ...همردیفی ... ، دقت کاهش نیابد. specifity) vs (precision

۶-۳ شبیه‌ساز

یکی از مراحل حساس و دشوار در تولید یک aligner , مرحله تست و آزمایش و ارزیابی نتایج است. برای این منظور امکان استفاده از خوانده‌های واقعی وجود ندارد به این علت که یک مکان قطعی برای هم‌ردیفی خوانده‌ها بر روی ژنوم وجود ندارد و نمی‌توان نظر قطعی در مورد مکان درست خوانده‌ها بر روی ژنوم داد. به همین منظور ابزارهایی برای شبیه‌سازی و تولید خواندهٔ مصنوعی از روی ژنوم وجود دارد (بهتره اینجا اسم چندتا ابزار مثل سم تولز اینا رو بیاریم). ما در ابتدا از این ابزارها برای تولید خوانده‌های مصنوعی استفاده کرده و آزمایش‌های اولیه را انجام دادیم, اما پس از پیش‌روی بیشتر, نیاز دیدیم که یک شبیه‌ساز تولید کنیم که با دریافت درصدهای متیلیشن, خطا و snip خوانده‌های مصنوعی تولید کند و همچنین درصد متیلیشن هر سیتوزین و گوانین را در یک فایل خروجی دهد. مشکلاتی که در دیگر شبیه‌سازها وجود داشت عدم توجه به جزایر CpG و عدم توانایی تولید خوانده به طور خاص از بعضی نواحی ژنوم بود . امروزه بخش عمده ای از پژوهشهای Methylation DNA برای کاهش هزینه ، به صورت RRBS انجام می‌شود . به این صورت که درصد بسیاری زیادی از خوانده‌ها از نواحی island CpG (بخشهایی از ژنوم که چگالی contex CpG ها زیاد است)که بخش کوچکی از ژنوم را تشکیل می‌دهند بوسیله افزودن آنزیم‌هایی (مانند (Mspl به دی‌ان‌ای آن را بخش‌بندی می‌کنند و سپس خوانده‌های با تراکم بالا از این قسمتها تولید و تست می‌شوند و شبیه‌سازهای موجود معمولا امکان شبیه‌سازی چنین خوانده‌هایی را فراهم نمی‌کنند. مشکل دیگر که ذکر شد همان عدم توجه به نواحی از ژنوم که به عنوان جزایر CpG تلقی می‌شوند است و شبیه‌ساز های موجود به متفاوت بودن درصد متیلاسیون CpG ها در این نواحی نسبت به خارج آنها توجهی ندارند و ما چون نیازمند آن بودیم که با خوانده‌هایی تست کنیم که این موضوع در ساخت آنها درنظر گرفته شده باشد ، شبیه‌ساز متناسب با این نیازمندی‌ها را پیاده‌سازی کردیم. سپس با استفاده از این شبیه‌ساز انواع مختلف خوانده را تولید کرده و در نهایت با هم‌ردیف کردن آنها و بدست آوردن درصدهای متیلیشن, نتایج را با خروجی شبیه‌ساز ، مورد مقایسه قرار دادیم.

۶-۴ محاسبه درصد متیلاسیون

برای بدست آوردن درصد متیلاسیون هر سیتوزین در samfile نهایی خروجی برنامه فوق ، که البته برای هر گونه فایل خروجی که در آن فرمت sam رعایت شده باشد قابل استفاده است ، برنامه دیگری نوشته شد. در این برنامه samfile خط به خط خوانده می شود و در ابتدا اگر خوانده‌های samfile ، خوانده های pcr هم باشند ، بر اساس flag مشخص کننده strand ای از ژنوم که با آن همردیف شده و تعداد تبدیل‌های آدنین به گوانین و سیتوزین به تیامین مشخص می‌شود که خوانده مربوط به کدام حالت زیر بوده است:

- تعداد تبدیلهای A به G بیشتر از C به T و flag همردیفی ۱۶
- تعداد تبدیلهای A به G بیشتر از C به T و flag همردیفی ۰
- تعداد تبدیلهای A به G کمتر از C به T و flag همردیفی ۱۶
- تعداد تبدیلهای A به G کمتر از C به T و flag همردیفی ۰
- تعداد تبدیلهای A به G حدودا برابر با C به T و flag همردیفی ۰ یا ۱۶

flag برابر با ۱۶ نشان‌دهنده این است که خوانده با رشته DNA مربوط به strand منفی همردیف شده است و flag صفر نشاندهنده هم‌ردیفی با strand مثبت . حالت اول و سوم مربوط به خوانده‌های pcr هستند و حالت دوم و چهارم مربوط به خوانده‌های غیر .pcr پس با این شمارش درصورتی که خوانده‌های pcr هم داشته باشیم ، می‌توانیم آنها را تمییز دهیم. مشکلی که باقی می‌ماند حالت پنجم است که در آن ما با یک خوانده مبهم روبرو هستیم که در این حالت هم می‌توان از آنها صرف نظر کرد و هم به صورت رندم یکی از دو حالت ممکن در نظر گرفت . پس از مشخص شدن نوع خوانده، سیتوزین‌های ژنوم مرجع با مقادیر متناظر در خوانده‌ها مقایسه می‌شوند و اگر در خوانده C دیده شود به تعداد خوانده هایی که در آنها C متیله بوده است و آن جایگاه در ژنوم را پوشش می‌دهند یک واحد اضافه می‌شود و اگر T دیده شود یک واحد به تعداد هایC غیر متیله اضافه می‌شود. در واقع پس از پایان این فرآیند برای تمامی خطوط samfile اطلاعاتی که ما به ازای هر سیتوزین در ژنوم مرجع خواهیم داشت عبارتند از:

- جایگاه سیتوزین نسبت به ابتدای کروموزوم

- ایstrand که خوانده با آن همردیف شده است
- تعداد خوانده هایی که این سیتوزین را پوشش داده و در آنها این C ، C بوده است.(تعداد خوانده های متیله)
- تعداد خوانده هایی که این سیتوزین را پوشش داده و در آنها این C ، بودهT است.(تعداد خوانده های غیرمتیله)

و با داشتن این مقادیر ، درصد متیلاسیون هر سیتوزین به صورت strand-specific بدست می‌آید.

فصل ۷

پیاده‌سازی

در اولین گام پیاده‌سازی لازم است که ژنوم‌های مورد نیاز (که در بخش solution مطرح گردید) از روی ژنوم اصلی ساخته شوند. از آنجایی که معمولا اندازهٔ ژنوم‌ها بسیار بزرگ است, این مرحله زمان زیادی می‌برد. لازم به ذکر است که این گام یک پیش‌پردازش است و برای هر ژنوم این مرحله تنها یک بار انجام می‌شود و پس از آن برای اجرای انواع alignment از همین ژنوم‌ها استفاده می‌شود. این بخش از برنامه به زبان C++ نوشته شده است. در گام بعدی, پارامترهایی به برنامهٔ آریانا اضافه گردید تا میان اجرای هم‌ردیفی در حالت بایسولفیت و غیر بایسولفیت تفاوت قائل شود. در حالتی که اجرای بایسولفیت مد نظر باشد, آریانا به ازای هر ژنومِ ساخته شده یک مرتبه هم‌ردیفی انجام می‌دهد. نکتهٔ قابل توجه آن است که برنامهٔ آریانا به گونه‌ای تغییر داده شده است که پیش‌پردازش‌ها و پس‌پردازش‌های مشترک میان هم‌ردیفی‌ها تنها یک بار صورت گیرد که هزینهٔ کمتری پرداخت شود. خروجی این مرحله ۵ سم‌فایل است که حاصل ۵ بار هم‌ردیفی است. در گام آخر باید از میاد هم‌ردیفی‌های صورت گرفته (۵ سم‌فایل) بهترین را انتخاب کنیم. بدیهی‌ترین روش برای انجام این مورد آن است که به ازای هر خوانده, تمامی سم‌فایل‌ها را بررسی کرده و بهترین هم‌ردیفی را انتخاب کنیم. واضح است که انجام این عمل به خودی خود امکان‌پذیر نیست, چرا که نمی‌توان ۵ فایل با ابعاد بسیار بزرگ را در حافظه نگه داشت و همچنین نمی‌توان به ازای هر خوانده تمامی فایل‌ها را یک بار جستجو کرد. به همین منظور در ابتدا سم‌فایل‌ها را بر اساس نام خوانده‌ها, توسط command linux مناسب مرتب می‌کنیم. سپس هر سطر از هر ۵ سم‌فایل را می‌خوانیم و از میان آن‌ها بهترین هم‌ردیفی را انتخاب کرده و نتیجه را مستقیما در خروجی چاپ می‌کنیم. در این حالت در هر لحظه تنها ۵ خوانده در حافظه وجود دارند و مشکل حافظه برطرف

می‌گردد. قابل توجه است که آریانا, علی‌رغم دیگر aligner ها, به ازای خوانده‌های ناموفق در هم‌ردیفی نیز سطری در سم‌فایل نهایی قرار می‌دهد. به همین دلیل, ترتیب خوانده‌ها پس از مرتب‌سازی در تمامی ۵ فایل یکسان است.

۷-۱ تابع پنالتی

تشخیص نواحی context CpG بر این اساس بود که هر سیتوزین در strand forward چک می‌شود که آیا بعد آن گوانین آمده است یا خیر و اگر آمده باشد یک ناحیه در نظر گرفته می‌شود. مکانهای جزایر CpG که تعداد آنها برای ژنوم انسان از O() را در حافظه به صورت مرتب‌شده نگه می‌داریم و برای چک کردن اینکه یک سیتوزین در جزیره CpG قرار دارد یا نه بر روی داده ساختاری که مکان شروع و پایان جزایر را نگه داشته شده ، جستجوی دودویی صورت می‌پذیرد. در تابع فعلی برای محاسبه پنالتی بازها ، اگر باز در خوانده و ژنوم هر دو C باشد و بعد از آن G آمده باشد و جزیره CpG پنالتی متوسط می‌دهیم چون معمولا سیتوزین‌های درون جزایر غیر متیله اند و در خارج جزیره پنالتی کم اختصاص می‌یابد (به این دلیل که در این حالت انتظار می‌رود خوانده متیله باشد پس سیتوزین ، بعد از بایسولفیت سیتوزین باقی می‌ماند) و در خارج context CpG پنالتی زیاد اختصاص می‌یابد(به این دلیل که در این حالت انتظار می‌رود خوانده متیله نباشد). اگر باز در خوانده T باشد و مقدار متناظر در ژنوم C باشد در حالت اول پنالتی صفر داریم ، در حالت دوم پنالتی کم اختصاص داده می‌شود(به این دلیل که در این حالت انتظار می‌رود خوانده متیله باشد پس سیتوزین ، بعد از بایسولفیت باید سیتوزین باقی بماند) و در خارج context CpG پنالتی کم اختصاص می‌یابد(به این دلیل که در این حالت انتظار می‌رود خوانده متیله نباشد). برای حالات درج و حذف نیز بیشترین پنالتی ممکن در نظر گرفته می‌شود چون درج و حذف به وضوح حالت نامطلوبی برای ما به حساب می‌آید و لازم است که بین این حالات و حالت تطابق تفاوت محسوسی قائل شویم. به طور کلی خوانده‌ها به صورت ترتیبی از فایل خوانده می‌شوند و بر اساس رشته سیگار آنها مشخص می‌شود که کدام بازهای آنها با حذف و اضافه و کدام بازها با تطبیق یا عدم تطبیق هم‌ردیف شده‌اند و سپس به روش توضیح داده شده پنالتی کلی یک خوانده محاسبه می‌شود.

۷-۲ محاسبه درصد متیلاسیون

به دلیل ابعاد بزرگ مساله و محدودیت حافظه ای که وجود دارد خوانده‌های همردیف شده را خط به خط از فایل ورودی گرفته و در یک بافر حلقوی به سایز طول بزرگترین خوانده ، جایگاه سیتوزین‌ها به همراه مقادیر توضیح داده شده نگه داشته می‌شوند و در صورت رسیدن به انتهای یک خوانده بافر به اندازه اختلاف جایگاه شروع خوانده بعدی با شروع خوانده فعلی خالی می‌شود. برای map سیتوزینهای ژنوم به بافر از یک تابع hash خطی ساده استفاده می‌شود.

فصل ۸

نتیجه‌گیری

در این فصل، ضمن جمع‌بندی نتایج جدید ارائه‌شده در پایان‌نامه، مسائل باز باقی‌مانده و همچنین پیش‌نهادهایی برای ادامه‌ی کار ارائه می‌شوند.

پیوست آ

مطالب تکمیلی

پیوست‌های خود را در صورت وجود می‌توانید در این قسمت قرار دهید.

واژه‌نامه

الف

ابتکاری ... heuristic
ارزش ... worth
ارضاپذیری ... satisfiability
استراتژی ... strategy
ائتلاف ... coalition

ب

بارگذاری ... loading
بازی ... game
برچسب ... label
برنامه‌ریزی خطی ... linear programming
برنامه‌ریزی صحیح ... integer programming
بسته‌بندی ... packing
بهترین پاسخ ... best response
بیشینه ... maximum

پ

پالت ... pallet
پایداری ... robustness
پشتیبان ... support
پوسته‌ی محدب ... convex hull
پوش بالایی ... upper envelope
پوششی ... covering

ت

تبدیل تصویری ... projective transformation
تعادل ... equlibrium
تعدیل ... relaxation
تقاطع ... intersection
تقسیم‌بندی ... partition
تکاملی ... evolutionary
توزیع‌شده ... distributed

ج

جست‌وجوی جامع ... brute-force
جست‌وجوی عمق‌اول ... Depth-First Search
جعبه ... bin

چ

چاله ..... sink

ح

حرکت ..... action

خ

خودخواهانه ..... selfish

خوشه ..... clique

د

دودویی ..... binary

دوگان ..... dual

دو ماتریسی ..... bimatrix

ر

رأس ..... vertex

رفتار ..... behaviour

رنگ‌آمیزی ..... coloring

ز

زمان‌بندی ..... scheduling

زیست‌شناسی ..... biology

س

ساختی ..... constructive

سود ..... pay off, utility

ش

شبه‌چندجمله‌ای ..... quasi-polynomial

شبه‌مقعر ..... quasi-concave

ص

صوری ..... formal

ع

عاقل ..... rational

عامل-محور ..... agent-based

عمل ..... action

غ

غائب ..... missing

غیرمتمرکز ..... decentralized

غیرمعمول ..... degenerate

ق

قابل انتقال ..... transferable

قاموسی ..... lexicographically

قوی ..... strong

ک

کمینه ..... minimum

م

مجموع زیرمجموعه‌ها ..... subset sum
مجموعه ..... set
محور ..... pivot
مختلط ..... mixed
مخفی ..... hidden
مستوی ..... affine
مسطح ..... planar
منطقی ..... reasonable
موازی ..... parallel

ن

نتیجه‌ی نهایی ..... outcome
نش ..... Nash
نقطه‌ثابت ..... fixed point
نگارخانه‌ی هنر ..... art gallery
نگهبان ..... gaurd
نمایه ..... profile
نوبتی ..... round-robin

و

وجه ..... facet

هـ

هزینه‌ی آشوب ..... price of anarchy (POA)
هزینه‌ی اجتماعی ..... social cost
هزینه‌ی پایداری ..... price of stability (POS)

ی

یال ..... edge
یکریختی ..... isomorphism

Abstract

We present a standard template for typesetting theses. The template is based on the XePersian package for the LaTeX typesetting system. This write-up shows a sample usage of this template.

Keywords: Thesis, Typesetting, Template, XePersian

Sharif University of Technology

Department of Computer Engineering

B.Sc. Thesis

DNA Bisulfite Sequencing by ARYANA

By:

Afsoon Afzal, Maryam Rabiee Hashemi

Supervisor:

Dr. Heydarnoori

Dr. Sharifi-Zarchi

January 2015