



دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی
گرایش مهندسی نرم‌افزار

عنوان:

هم‌ردیفی خوانده‌های با ای‌سولفیت توسط آریانا

نگارش:

افسون افضل، مریم ربیعی هاشمی

استاد راهنما:

دکتر حیدرنوری

دکتر شریفی زارچی

۱۳۹۳ بهمن

مَلِكُ الْأَفْلَامِ

به نام خدا
دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی

عنوان: هم‌ردیفی خوانده‌های بایسولفیت توسط آریانا
نگارش: افسون افضل، مریم ریبعی هاشمی

سپاس

از استاد بزرگوارم که با کمک‌ها و راهنمایی‌های بی‌دriegشان، بنده را در انجام این پژوهه یاری داده‌اند،
تشکر و قدردانی می‌کنم. همچنین از دوست عزیزم که در به دست آوردن نتایج این مقاله همکاری داشتند،
صمیمانه سپاس‌گزارم.

چکیده

نگارش پایان نامه علاوه بر بخش پژوهش و آماده سازی محتوا، مستلزم رعایت نکات فنی و نگارشی دقیقی است که در تهیه یک پایان نامه موفق بسیار کلیدی و مؤثر است. از آن جایی که بسیاری از نکات فنی مانند قالب کلی صفحات، شکل و اندازه قلم، صفحات عنوان و غیره در تهیه پایان نامه ها یکسان است، با استفاده از نرم افزار حروف چینی زی تک و افزونه زی پرشین یک قالب استاندارد برای تهیه پایان نامه ها ارائه گردیده است. این قالب می تواند برای تهیه پایان نامه های کارشناسی و کارشناسی ارشد و نیز رساله دکتری مورد استفاده قرار گیرد. این نوشتار به طور مختصر نحوه استفاده از این قالب را نشان می دهد.

کلید واژه ها: پایان نامه، حروف چینی، قالب، زی پرشین

فهرست مطالب

۱۰	۱	مقدمه
۱۱	۱-۱	تعريف مسئله
۱۱	۲-۱	اهمیت موضوع
۱۲	۳-۱	اهداف تحقیق
۱۲	۴-۱	ساختار پایان نامه
۱۳	۲	مفاهیم اولیه
۱۳	۱-۲	رشته‌ی DNA
۱۴	۲-۲	همانندسازی DNA
۱۴	۳-۲	ژنوم
۱۵	۴-۲	خوانده
۱۶	۵-۲	متیلاسیون DNA
۱۶	۶-۲	خوانده‌های بایسولفیت شده
۱۷	۷-۲	File Sam
۱۷	۸-۲	رشته‌ی CIGAR
۱۷	۹-۲	توالی‌یابی نسل بعد
۱۷	۱۰-۲	Alignment Sequence

۱۸	Sequencing DNA	۱۱-۲
۱۸	آریانا	۱۲-۲
۱۹	کارهای پیشین	۳
۱۹	الگوریتم های ابتدایی	۱-۳
۲۰	ابزار بیسمارک	۲-۳
۲۰	BSmap	۳-۳
۲۱	کاستی های روش های پیشین	۴-۳
۲۲	راه حل	۴
۲۲	الگوریتم	۱-۴
۲۵	انتخاب بهترین هم ردیفی	۲-۴
۲۶	شبیه ساز	۳-۴
۲۷	محاسبه درصد متیلاسیون	۴-۴
۲۹	پیاده سازی	۵
۳۰	تابع پنالتی	۱-۵
۳۱	محاسبه درصد متیلاسیون	۲-۵
۳۲	نتیجه گیری	۶
۳۳	مطالب تکمیلی	آ

فهرست شکل‌ها

۱۴	۱-۲ ساختار <i>DNA</i>
۱۵	۲-۲ همانندسازی <i>DNA</i>
۱۶	۳-۲ متیلاسیون <i>DNA</i>
۲۳	۱-۴ حالات قابل قبول نگاشت C به T

∧

فهرست جداول‌ها

فصل ۱

مقدمه

متیلاسیون^۱ سیتوزین^۲ از بسیاری از جهات بر بیولوژی انسان تأثیرگذار است از جمله رشد جنینی و 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 نمونه 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 نشان‌دهنده متبیله بودن و نگاشت T به C نشان‌دهنده غیر متبیله بودن است. روش‌ها و الگوریتم‌های گوناگونی برای توالی‌یابی خوانده‌های بایسولفیت شده ارائه شده‌اند و بر اساس آنها ابزارهای توسعه یافته‌اند. با این حال ضعف این ابزارها در کمی دقیق و همچنین عدم درنظرگیری خواص زیستی است. به همین دلیل ما تلاش نمودیم که با توسعه ابزار آریانا، که یک

Methylation^۱

Cytosine^۲

transcription^۳

chromatin structure^۴

DNA repair^۵

DNA repair^۶

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

۱-۱ تعریف مسئله

در این پژوهه ما سعی داشتیم که یک aligner برای خوانده‌های treated bisulfite بنویسیم. هدف مسئله، نگاشت یک تعداد از رشته‌های متشكل از حروف G,C,T,A بر روی یک ژنوم خاص می‌باشد که تمامی این رشته‌ها از یک ژنوم بدست آمده‌اند ولی کاملاً مشابه آن نیستند و تغییراتی در آنها صورت پذیرفته است. هدف یافتن جایگاه این رشته‌ها در ژنوم با دقت حداکثر و درنظر داشتن محدودیتها زمانی و حافظه‌ای با توجه به ابعاد مسئله است. در این پژوهه 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 پژوهه بر آن بود که راه حلی 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.

۱-۲ اهمیت موضوع

مسئله‌ی مسیریابی وسائل 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 وسائل نقلیه در [؟، ؟، ؟] بیان شده است.

۱-۳ اهداف تحقیق

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

۱-۴ 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 حداقل اندازه‌ی درخت‌ها می‌پردازیم. فصل پنجم به نتیجه‌گیری و پیش‌نهادهایی برای کارهای آتی خواهد پرداخت.

فصل ۲

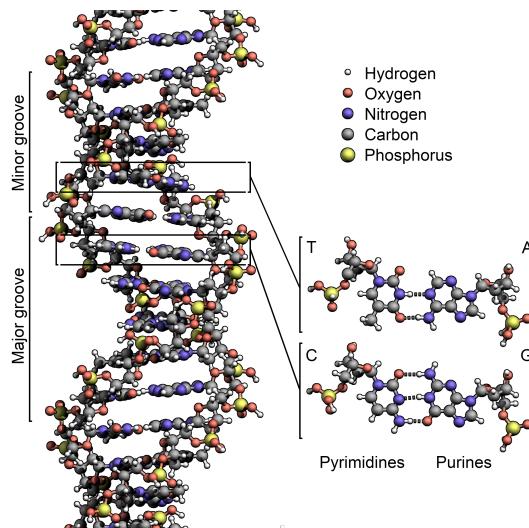
مفاهیم اولیه

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

۱-۲ 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 فردیک 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 نوکئوتید از یک باز نوکلئوتیدی شامل نیتروژن تشکیل شده است که این بازها آدنین (A) ، تیامین (T) ، سیتوزین (C) و گوانین (G) که در زیر شکل ساختار آنها مشاهده می‌شود:

نوکلئوتیدها در یک زنجیره توسط پیوندهای کوالانسی بین قند یک نوکلئوتید و فسفات نوکلئوتید 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 و همینطور سیتوزین و گوانین رو بروی هم قرار بگیرند.



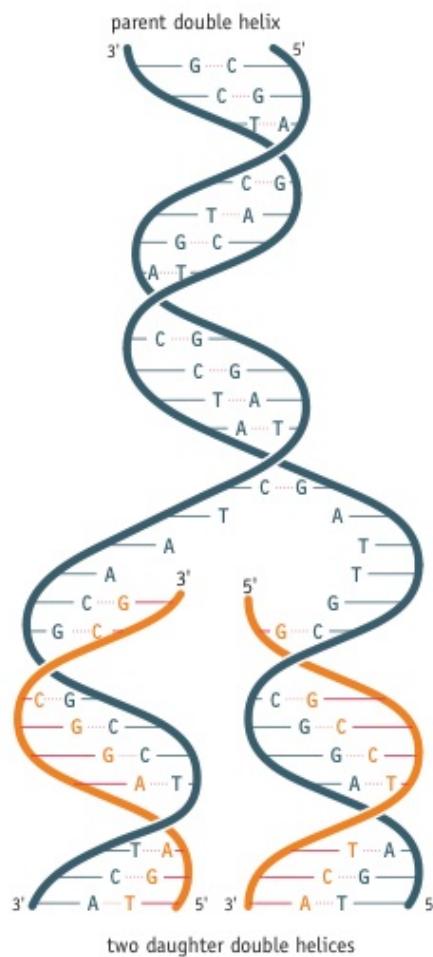
شکل ۱-۲: ساختار DNA

۲-۲ همانندسازی DNA

برای اینکه وراثت امکان‌پذیر باشد، ژنها باید توانایی همانندسازی داشته باشند. ژن‌ها 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 همانندسازی، 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 روی‌روی هر باز، باز مکمل آن قرار می‌گیرد، به این ترتیب با استفاده از نوکلئوتیدهای آزاد که در سیتوپلاسم وجود دارند، در مقابل T باز A، در مقابل C باز G قرار می‌گیرد و در آخر دو کپی کاملاً مشابه والد ساخته می‌شود.

۳-۲ ژنوم

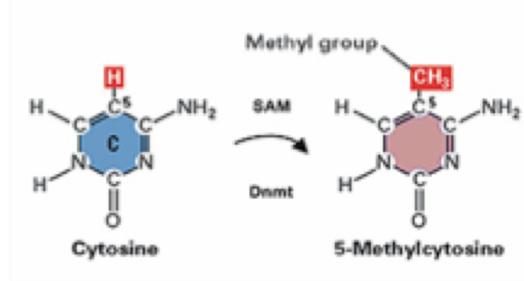
اطلاعات ژنتیکی حمل شده توسط DNA در توالی و ترتیب خطی DNA به واحدهای عملکردی مجزایی به نام ژنها تقسیم شده است که به طور شاخص حدود ۵۰۰۰۰ تا ۱۰۰۰۰۰ نوکلئوتید طول دارند. به مجموعه این ژن‌ها ژنوم گفته می‌شود (Waterman, ۱۹۹۵).



شکل ۲-۲: همانندسازی DNA

۴-۲ خوانده

مولکول دی ان ای 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 هر کدام از آنها جداگانه توالی‌یابی شوند. به این قطعات read (خوانده) می‌گویند.



شکل ۳-۲: متیلاسیون DNA

۵-۲ متیلاسیون DNA

متیلاسیون DNA شامل اضافه شدن یک گروه متیل به انتهای کربن سیتوزین، C5، توسط DNA methyltransferases می‌باشد. به سیتوزینی که گروه متیل به آن اضافه شده است متیل سیتوزین گفته می‌شود (شکل ۵-۲). (Krueger, ۲۰۱۲). (<http://www.ncbi.nlm.nih.gov/pubmed/22290186>)

۶-۲ خوانده‌های بايسولفیت شده

شامل خوانده‌هایی است که در محلول سدیم بايسولفیت 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 آمده است. بعد از یک سیتوزین در یک رشته، هم می‌تواند A، T، و یا C بیاید. بنابراین تمام سیتوزین‌ها در CpG context نمی‌باشند ولی تمام ها CpG به طور مشخص در G ها هستند (Gardiner, ۱۹۸۷).

۲. CpG islands: تعريف کاربردی از CpG islands باز بر می‌گردد که درصد تعداد ها CG بیشتر از ۵۰٪ باشد و نرخ مشاهده شده مورد انتظار G بیشتر از ۶۰٪ باشد. منظور از نرخ مشاهده شده مورد انتظار عددی است که از فرمول (۱-۲) بدست می‌آید (Gardiner, ۱۹۸۷).

File Sam ۷-۲

islands: CpG تعريف کاربردی از CpG islands ها به یک ناحیه به همراه حداقل ۲۰۰ جفت باز برمی گردد که درصد تعداد ها CG باشد و نرخ مشاهده شده مورد انتظار CpG بیشتر از ۵۰٪ است. منظور از نرخ مشاهده شده مورد انتظار عددی است که از فرمول $(1 - \frac{1}{2^L})$ بدست می آید (Gardiner, ۱۹۸۷).

CIGAR رشته‌ی ۸-۲

رشته‌ای است برای مشخص کردن اینکه کدام باز خوانده با کدام باز ژنوم مرجع هم‌ردیف شده است. این رشته از حروف m, d, i 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 جفت‌های عدد و کاراکتر به صورت پشت سر هم ظاهر می‌شوند.

توالی‌یابی نسل بعد ۹-۲

تکنولوژی NGS (Next Generation Sequencing) Technology مجموعه ای از تکنیک‌های جدید برای خواندن توالی DNA می‌باشد که از نظر دقیق 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).

Alignment Sequence ۱۰-۲

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

Sequencing DNA ۱۱-۲

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 sequencing گویند. متدها و تکنولوژی‌های مختلفی برای یافتن توالی DNA استفاده شده است (Waterman, ۱۹۹۵).

۱۲-۲ آریانا

آریانا یک برنامه هم‌ردیفساز است که از الگوریتم Wheeler Burrows یا 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 ۳

کارهای پیشین

روش‌های بسیاری برای هم‌ردیفی خوانده‌های با ایسلوفیت، 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 سیتوزین‌ژنوم 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 ungapped و یا gapped را بسته به برنامهٔ مورد استفاده، پیاده‌سازی کرد.

۱-۳ الگوریتم‌های ابتدایی

برنامه CokusAlignment از روشهای نگاشت 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 تقریباً کوچک اجرا

می شود. لازم به ذکر است که می توان با بهینه سازی برای پروژه های مختلف این سرعت را بهبود بخشید اما در بسیاری از پروژه ها چنین کاری امکان پذیر نیست. همچنین این روش برای پیاده سازی نگاشت های pair-end و gapped نیز مناسب نیست. از نظر عملی به دلیل کمبود سرعت و عملکرد، این روش قابل استفاده نیست.

۲-۳ ابزار بیسمارک

هدف ابزار Bismark، یافتن یک تطابق منحصر به فرد با چهار بار اجرای پردازش ها به صورت همزمان می باشد. در ابتدا های read 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 نوع های پیش تغییر یافته از ژنوم رفرنس نگاشت می شوند که این عمل با استفاده از نرم افزار نگاشت Bowtie و به صورت چهار نمونه موازی صورت می گیرد. این نگاشت ها، Bismark را قادر می سازد تا به صورت مشخص، رشته اصلی بایسولفیت را مشخص 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 می باشد. این اطلاعات را می توان به عنوان پس پردازش در نظر گرفت. همچنین Bismark خروجی متیلاسیون را در بین ناحیه های مختلف CPG, CHG, CHH در نظر می گیرد و این ناحیه ها را از هم جدا در نظر می گیرد

۳-۳ ابزار BSmap

BSmap از این قضیه اصلی که تمام جایگاه های سیتوزین که تغییرات نا مترافقن C/T (سیتوزین به تیامین) در آنها رخ می دهدن ، شناخته شده هستند، برای راهنمایی در هم ردیفی های read بایسولفیت شده استفاده می کند. BSMAP تیامین های موجود در خوانده های بایسولفیت شده را به عنوان سیتوزین mask می نماید. (بر عکس تغییرات بایسولفیت) که این تغییر را فقط در جایگاه های سیتوزین ژنوم اصلی انجام می دهد در حالیکه کل تیامین های دیگر را در خوانده های بایسولفیت شده بدون تغییر نگه می دارد. بعد از انجام این تغییرات ، خوانده های بایسولفیت شده را به طور مستقیم بر روی ژنوم نگاشت می

نماید. (Chen, ۲۰۱۰) علاوه بر موارد ذکر شده، BSMAP براساس الگوریتم کاراتر table HASH seeding کار می‌کند به این صورت که ژنوم مرجع را برای تمام های k-mer ممکن شاخص گذاری می نماید که هر کدام از آنها seed خوانده می‌شود. برای یافتن یک نگاشت، تنها جایگاه‌هایی به (seed) طور کامل با بخشی از خوانده منطبق شده اند جست و جو می‌شوند. با جست و جو در جدول، *seed* اکثریت جایگاه‌های غیرنگاشت شده دور ریخته می‌شوند و کارایی جست و جو به صورت قابل ملاحظه ای افزایش می‌یابد.(Chen, ۲۰۱۰).

۴-۳ کاستی‌های روش‌های پیشین

این روش خواهد های 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 های mismatch مجاز در نرم‌افزار مورد استفاده است. این تعداد می‌تواند توسط های mismatch واقعی (SNP) بیشتر کاسته شود و نتیجه را بیش از قبل محدود کند.

فصل ۴

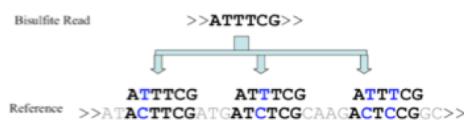
راه حل

۱-۴ الگوریتم

روشی که برای هم‌ردیفی خوانده‌های بایسولفیت در aligner های امروزی مانند Brat و Bismark 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 دقیق کافی با توجه به اهمیت پیدا کردن نقاط مตیله شده نکرده‌اند و در این‌ها برای بالا بردن دقیقی و پیدا کردن مکان درست منتظر 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 نباشد حالت اول پیش می‌آید، اگر متیله نباشد حالت دوم و حالت سوم نیز هم‌ردیفی T با T است و مشکلی ندارد.



شکل ۱-۴: حالات قابل قبول نگاشت C به T

حالت آخر در هیچ شرایطی امکان‌پذیر نیست و این هم‌ردیفی قابل قبول نمی‌تواند باشد ولی با توجه 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. رشته‌های crick Watson که در واکنش با بایسولفیت تغییر یافته‌اند دیگر مکمل یکدیگر نمی‌باشند زیرا تغییرات بایسولفیت فقط در سیتوزین‌های غیرمتیله رخ می‌دهد.

نگاشت سیتوزین‌ها به تیامین‌ها به صورت نامتقارن انجام می‌شود. تیامین موجود در های read بایسولفیت شده هم می‌تواند به سیتوزین و هم می‌تواند به تیامین در رفرنس ژنوم نگاشت شود اما این امر به صورت عکس امکان‌پذیر نمی‌باشد. این مشکل نه تنها فضای جست و جو را افزایش می‌دهد بلکه پروسه نگاشت را نیز پیچیده تر می‌نماید. مانند مثالی که در ادامه آورده شده است (شکل ۴-۲). 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 براین 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. سیتوزین‌های CpG context با احتمال ۹۰٪ متیله هستند و سیتوزین‌های 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 نوکلئوتیدهای CpG آنها چگالی بالایی دارند و سیتوزینهای این های CpG 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

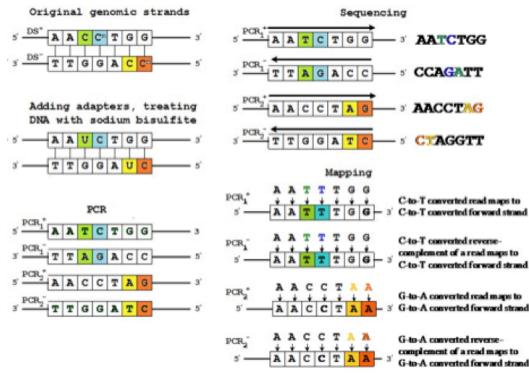
نظر گرفت ، تولید می کنیم . به جای آنکه تمامی سیتوزین ها را به تیامین تبدیل کنیم به صورت انتخابی و با توجه به مکانی که آن سیتوزین در آن قرار گرفته است این کار را انجام می دهیم ؛ به این صورت که فرض را بر این می گذاریم که سیتوزین هایی که در نواحی CpG context قرار دارند متیله هستند بنابراین پس از بایسولفیت شدن تغییری نمی کنند و سیتوزین هایی که در خارج از این نواحی هستند با احتمال خوبی متیله نیستند و پس از بایسولفیت شدن به تیامین تبدیل می شوند. همچنین با توجه به اینکه سیتوزین های درون جزایر CpG با احتمال بیشتری متیله نیستند ، سیتوزین هایی که درون نواحی CpG و context CpG تغییر باقی مانندند. ژنومی که با تغییرات فوق بدست می آید برای هم ردیفی تمامی خوانده های بایسولفیت کافی نیست . علت آن این است که در بدست آوردن خوانده ها حتی با روش های Next sequencing خطاب وجود دارد و ممکن است که در خوانده ای تیامین به اشتباه سیتوزین گرفته شود (چه فعلی جز خوانده شود؟) و در فرآیند هم ردیفی آن خوانده مشکل بوجود آورد . دلیل دیگر این است که در سه حالت فوق ما فرض کردیم که تمامی سیتوزین های خوانده های بیرون CpG و یا درون جزایر CpG متیله نیستند و تمامی خوانده های درون ناحیه و خارج از جزایر متیله اند ولی در واقعیت ما این قطعیت را نداریم . برای جبران این تفاوت ما علاوه بر ژنوم فوق ژنوم دیگری تولید می کنیم و در آن تمامی سیتوزین ها را به تیامین تبدیل می نماییم . بدین صورت خوانده هایی که با ژنوم قبل نتوانند هم ردیف شوند با این ژنوم خواهند شد. برای اینکه هم ردیفی با خواص زیستی تطابق بیشتری داشته باشد در انتخاب هم ردیفی نهایی به ژنوم قبل اولویت داده می شود. علاوه بر این دو ژنوم نیاز این دیده می شود که برای اینکه خوانده های بیشتری را بتوانیم هم ردیف کنیم ، خوانده ها با ژنوم اصلی نیز هم ردیف شوند . دلیل آن هم این است که ممکن است نواحی ای وجود داشته باشند که سیتوزین ها علی رغم اینکه CpG نیستند کاملاً متیله باشند و ظاهر این نواحی مشابه ژنوم اصلی باشد. دلیل بعدی این موضوع این است که treatment bisulfite روی همه خوانده ها کامل عمل نمی کند و ممکن است یک سیتوزین غیر متیله به تیامین تبدیل نشود . برای پوشش دادن این حالات نیز بهتر است که از ژنوم اصلی برای هم ردیفی استفاده شود. علاوه بر ژنوم های 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 خوانده ای وجود داشته باشد که مطابق انتظار ما نباشد. تلاش ما بر این بود که تا جای ممکن تمامی حالات ممکن برای خوانده ها را مدنظر قرار بدهیم. لازم به ذکر است که پس از انجام آزمایش های فراوان با خوانده های شبیه سازی شده و

واقعی و نتیجه لازم را از این ژنوم اضافه را دریافت نکردیم و وجود آن را بی مورد 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 سیتوزین ها در خوانده نیز به تیامین تبدیل شده و سپس هم ردیفی انجام می شود اما در راه حلی که ما ارائه کردہ ایم چنین عملی بی مورد به نظر می رسد. باید توجه داشت که اکثر خوانده هایی که هم ردیفی ناموفقی با ژنوم ساخته شده در قسمت قبل داشته اند، خوانده هایی هستند که سیتوزین های آنها متیله نبوده و به تیامین تبدیل شده اند ولی ما در ژنوم ساخته شده به اشتباه آنها را سیتوزین نگه داشته ایم. در این حالت دیگر نیازی به تغییر در خوانده نیست و باید بدون تغییر به ژنومی که تمامی سیتوزین های آن به تیامین تبدیل شده است، هم ردیف گردد.

یکی از موارد بسیار مهم در زمینه هم ردیفی خوانده های بایسولفیت شده، توانمندی aligner در هم ردیفی خوانده های PCR شده است. می توان نشان داد که در این حالت، ژنوم های مطرح شده کافی نیستند و خوانده های مورد نظر را پوشش نمی دهند. همانطور که در شکل زیر پیداست، در این نوع، خوانده ها هم از رشتہ مثبت و هم از رشتہ منفی جمع آوری می گردند و در دستگاه PCR تکثیر می شوند. سیتوزین های غیر متیله در این خوانده ها، پس از بایسولفیت شدن به تیامین تبدیل می شوند. به همین دلیل، اگر خوانده ای از رشتہ منفی برداشته شده باشد، پس از بایسولفیت شدن معادل ان خواهد بود که در معکوس این خوانده گوانین های غیر متیله به آدنین تبدیل شده باشند. بدیهی است که این مورد را نمی توان با ژنوم های پیشتر مطرح شده پوشش داد. به همین دلیل ژنوم های ساخته می شود که در آنها در رشتہ مثبت، گوانین ها به آدنین تبدیل می گردد. البته ما همچنان خواص زیستی را در این مورد نیز مورد توجه قرار داده ایم و برای ساخت این ژنوم مناطق CG و مناطق جزیره CpG را مد نظر قرار می دهیم.
(عکس، شکل)

۲-۴ انتخاب بهترین هم ردیفی

به منظور انتخاب بهترین هم ردیفی برای هر خوانده از میان هم ردیفی هایی مختلفی که می تواند با ژنوم های تولید شده داشته باشد ، به محاسبه پنالتی می پردازیم . 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 می گیریم . پنالتی یک خوانده در حال حاضر جمع پنالتی های بازه ای آن می باشد ولی برنامه نوشته شده به این صورت است که قابلیت تغییر تابع محاسبه



شکل ۲-۴: تبدیل PCR

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 خوانده‌ها بدست می‌آید سقف در نظر بگیریم تا به بهای ... هم‌ردیفی ...، دقت کاهش نیابد. specificity) vs (precision)

۳-۴ 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 خوانده 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، خط 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 عمده‌ای از پژوهش‌های Methylation DNA برای کاهش هزینه، به صورت RRBS انجام می‌شود. به این صورت که درصد بسیاری زیادی از خوانده‌ها از نواحی CpG island (بخشهایی از ژنوم که چگالی CpG ها زیاد است) که بخش کوچکی از

ژنوم را تشکیل می‌دهند بوسیله افزودن آنزیم‌هایی (مانند $MspI$) به دی‌ان‌ای آن را بخش‌بندی می‌کنند و سپس خوانده‌های با تراکم بالا از این قسمتها تولید و تست می‌شوند و شبیه‌سازهای موجود معمولاً 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 شبیه‌سازهای 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.

۴-۴ 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 می‌شود و در ابتدا اگر خوانده‌های samfile ، خوانده‌ای 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 حدوداً 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 خوانده‌هایی 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 های C 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
- ای strand که خوانده با آن هم ردیف شده است
- تعداد خوانده‌هایی که این سیتوزین را پوشش داده و در آنها این C ، C بوده است.(تعداد خوانده های متیله)
- تعداد خوانده‌هایی که این سیتوزین را پوشش داده و در آنها این C ، بوده T است.(تعداد خوانده های غیرمتیله)

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

فصل ۵

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 معمولاً اندازه 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 برنامه آریانا اضافه گردید تا میان اجرای هم‌ردیفی در حالت بایسولفیت و غیر بایسولفیت تفاوت قائل شود. در حالتی که اجرای بایسولفیت مدنظر باشد، آریانا به ازای هر ژنوم ساخته شده یک مرتبه هم‌ردیفی انجام می‌دهد. نکته قابل توجه آن است که برنامه آریانا به گونه‌ای تغییر داده شده است که پیش‌پردازش‌ها و پس‌پردازش‌های مشترک میان هم‌ردیفی‌ها تنها یک بار صورت گیرد که هزینه 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. بدیهی ترین 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 مستقیماً 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 سطیری در سمت فایل نهایی قرار می‌دهد. به همین دلیل، ترتیب خوانده‌ها پس از مرتب‌سازی در تمامی ۵ فایل یکسان است.

۱-۱ تابع پنالتی

تشخیص نواحی 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 مرتب شده 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 معمولاً سیتوزین‌های درون جزایر غیر مตیله 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.

۲-۵ 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 زنوم به بافر از یک تابع hash خطی ساده استفاده می‌شود.

فصل ۶

نتیجه‌گیری

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

آ پیوست

مطالب تکمیلی

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

واژه‌نامه

support	پشتیبان	الف
convex hull.....	پوسته‌ی محدب	ابتکاری
upper envelope.....	پوش بالایی	ارزش
covering	پوششی	ارض‌آپذیری
		استراتژی
		ائتلاف
		ت
projective transformation.....	تبديل تصویری	
equilibrium.....	تعادل	ب
relaxation.....	تعديل	بارگذاری
intersection.....	تقاطع	بازی
partition.....	تقسیم‌بندی	برچسب
evolutionary	تکاملی	برنامه‌ریزی خطی
distributed	توزیع شده	برنامه‌ریزی صحیح
		بسته‌بندی
		بهترین پاسخ
		بیشینه
		ج
brute-force	جست‌وجوی جامع	
Depth-First Search.....	جست‌وجوی عمق‌اول	
bin	جعبه	پ
		پالت
		پایداری

		ج
ش	sink	چاله
quasi-polynomial	شبیه‌چندجمله‌ای	
quasi-concave	شبیه‌مقعر	ح
	action	حرکت
ص		
formal	صوری	خ
	selfish	خودخواهانه
ع	clique	خوش
rational	عاقل	
agent-based	عامل-محور	د
action	عمل	دودویی
	binary	دوگان
غ	dual	دو ماتریسی
missing	غائب	
decentralized	غیر مرکز	ر
degenerate	غیر معمول	رأس
	vertex	رفتار
ق	behaviour	رنگ‌آمیزی
transferable	قابل انتقال	
lexicographically	قاموسی	ز
strong	قوی	زمان‌بندی
	scheduling	زیست‌شناسی
ک	biology	
minimum	کمینه	س
	constructive	ساختی
	pay off, utility	سود

gaurd	نگهبان	نگهبان	م
profile	نمایه	subset sum	مجموع زیرمجموعه‌ها
round-robin	نوبتی	set	مجموعه
		pivot	محور
	و	mixed	مختلط
		hidden	مخفى
		affine	مستوى
	ه	planar	مسطح
price of anarchy (POA)	هزینه‌ی آشوب	reasonable	منطقی
social cost	هزینه‌ی اجتماعی	parallel	موازی
price of stability (POS)	هزینه‌ی پایداری		
			ن
	ی	outcome	نتیجه‌ی نهایی
edge	یال	Nash	نش
isomorphism	یکریختی	fixed point	نقطه ثابت
		art gallery	نگارخانه‌ی هنر

Abstract

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

Keywords: Thesis, Typesetting, Template, X_EPersian



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