



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 بزرگوارمان جناب آقای دکتر شریفی زارچی که با کمک‌ها و راهنمایی‌های بی‌دriegشان، ما را در راه انجام این پروژه یاری داده‌اند، تشکر و قدردانی می‌نماییم. همچنین از جناب آقای دکتر چیت‌ساز برای فراهم کردن محیط تست بر روی رایانه‌های قدرتمند قدردانی می‌نماییم. از جناب آقای دکتر حیدرنوری نیز برای حمایت و پشتیبانی سپاس‌گزاریم.  
از پژوهشگاه رویان و تمام کسانی که ما را در انجام این پروژه یاری نمودند، تشکر می‌نماییم.

## چکیده

متیلاسیون سیتوزین به علت تأثیرگذاری بالا بر روی فرایندها زیستی مختلف از اهمیت شایانی برخوردار است. نمایان شدن متیلاسیون سیتوزین‌ها با افزودن محلول سدیم بایسولفیت به خوانده‌های بدست آمده از ژنوم، صورت می‌گیرد و هم‌ردیفی بادقت و کارای چنین خوانده‌هایی نقش اساسی در محاسبه درصد متیلاسیون دارد.

روش‌های مختلفی تا به حال برای این مسئله مطرح شده است که تمرکز هرکدام بریکی از زمینه‌های سرعت، دقت و کارایی بوده است. در این پایان‌نامه، راه‌کار جدیدی برای این مسئله با گسترش آریانا، که ابزاری برای هم‌ردیفی خوانده‌ها است، ارائه خواهد شد که بر افزایش دقت هم‌ردیفی خوانده‌های بایسولفیت و استفاده از خواص زیستی از جمله خواص سیتوزین‌ها در جزایر و نواحی  $CpG$ ، تأکید می‌کند.

با استفاده از این راه‌کار دقت هم‌ردیفی و توانایی هم‌ردیفسازی خوانده‌های بایسولفیت، به خصوص در مواردی که خوانده‌ها از نواحی ارزشمند ژنوم که همان CpG context‌ها هستند افزایش یافته است و درصد متیلاسیون سیتوزین‌های خوانده‌های شبیه‌سازی شده با اختلاف قابل قبول با مقادیر حقیقی بدست می‌آید.

کلیدواژه‌ها: بایسولفیت، هم‌ردیفی، متیلاسیون، جزایر  $CpG$ ، آریانا

# فهرست مطالب

۱۰	۱	مقدمه
۱۱	۱-۱	تعريف مسئله
۱۱	۲-۱	اهداف تحقیق
۱۲	۳-۱	ساختار پایان نامه
۱۳	۴	مفاهیم اولیه
۱۳	۱-۲	رشته‌ی DNA
۱۴	۲-۲	هماندسازی DNA
۱۵	۳-۲	ثنوم
۱۶	۴-۲	خوانده
۱۶	۵-۲	متیلاسیون DNA
۱۶	۶-۲	خوانده‌های بایسولفیت شده
۱۷	۷-۲	فایل Sam
۱۷	۸-۲	رشته‌ی سیگار
۱۸	۹-۲	همردیفسازی توالی
۱۸	۱۰-۲	آریانا

۱۹	۳	کارهای پیشین
۱۹	۱-۳	الگوریتم های ابتدایی
۲۰	۲-۳	ابزار <i>Bismark</i>
۲۰	۳-۳	ابزار <i>BSmap</i>
۲۱	۴-۳	ابزار <i>BS - Seeker</i>
۲۱	۵-۳	کاستی های روش های پیشین
۲۲	۴	راه حل
۲۲	۱-۴	الگوریتم
۲۶	۲-۴	انتخاب بهترین هم ردیفی
۲۶	۳-۴	شبیه ساز
۲۷	۴-۴	محاسبه درصد متیلاسیون
۲۹	۵	پیاده سازی
۳۰	۱-۵	تابع جریمه
۳۱	۲-۵	محاسبه درصد متیلاسیون
۳۲	۶	نتایج
۳۲	۱-۶	مقایسه زمان اجرا
۳۳	۲-۶	مقایسه توانایی هم ردیف سازی و دقت
۳۴	۳-۶	مقایسه توانایی هم ردیف سازی خوانده های <i>PCR</i>
۳۴	۴-۶	بررسی دقت محاسبه درصد متیلاسیون
۳۶	۵-۶	کارهای آینده

# 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
۲۵	.....	۲-۴ تبدیل <i>PCR</i>
۳۳	.....	۱-۶ مقایسه زمان اجرا
۳۴	.....	۲-۶ مقایسه توانایی هم‌ردیف‌سازی
۳۵	.....	۳-۶ مقایسه دقت هم‌ردیف‌سازی
۳۶	.....	۴-۶ مقایسه دقت و توانایی هم‌ردیف‌سازی
۳۷	.....	۵-۶ مقایسه دقت در خوانده‌های <i>PCR</i>
۳۷	.....	۶-۶ درصد متیلاسیون واقعی و به دست آمده
۳۸	.....	۷-۶ فراوانی اختلاف درصد متیلاسیون

## فهرست جداول‌ها

۱-۴ میزان نگاشت به هر ژنوم در رشتۀ مثبت ..... ۲۵

# فصل ۱

## مقدمه

متیلاسیون<sup>۱</sup> سیتوزین<sup>۲</sup> از بسیاری از جهات از جمله رشد جنینی، رونویسی<sup>۳</sup> و ساختار کروماتین<sup>۴</sup> بر بیولوژی انسان تأثیرگذار است. این مورد در گیاهان نیز به همان اندازه، در مواردی چون رونویسی، ترمیم DNA<sup>۵</sup> و تفاوت سلولی<sup>۶</sup>، اهمیت دارد. نکته قابل توجه آن است که متیلاسیون DNA<sup>۷</sup> در انعطاف و حافظه سیستم عصبی مؤثر است و همچنین متیلاسیون غیر طبیعی DNA عامل بسیاری از بیماری‌ها از جمله آלצהایمر و سرطان است. روش‌های مختلف درمانی سرطان در حال توسعه هستند که با هدف اصلاح الگوی متیلاسیون عمل می‌کنند.

روش استاندارد اندازه‌گیری متیلاسیون DNA، افزودن محلول سدیم بایسولفیت<sup>۸</sup> به نمونه برداشت شده است که سیتوزین‌های غیرمتیله<sup>۹</sup> را به یوراسیل<sup>۱۰</sup> (که پس از تکثیر به تیامین<sup>۱۱</sup> تبدیل می‌شود) تبدیل

Methylation<sup>۱</sup>

Cytosine<sup>۲</sup>

Transcription<sup>۳</sup>

Chromatin Structure<sup>۴</sup>

DNA repair<sup>۵</sup>

Cell Differentiation<sup>۶</sup>

Deoxyribonucleic Acid<sup>۷</sup>

Sodium Bisulfite<sup>۸</sup>

Unmethylated<sup>۹</sup>

Uracil<sup>۱۰</sup>

Thymine<sup>۱۱</sup>

می‌کند. پس از آن  $DNA$ ، توالی‌یابی<sup>۱۲</sup> می‌گردد و با ژنوم مرجع<sup>۱۳</sup> مقایسه می‌شود به طوری که نگاشت  $C$  به  $C$  نشان‌دهنده متیله بودن و نگاشت  $T$  به  $C$  نشان‌دهنده غیرمتیله بودن است [۱]. روش‌ها و الگوریتم‌های گوناگونی برای توالی‌یابی خوانده‌های بایسولفیت‌شده ارائه شده‌اند و بر اساس آنها ابزارهای توسعه یافته‌اند. با این حال ضعف این ابزارها در دقت پایین و همچنین عدم درنظرگیری خواص زیستی است. به همین دلیل ما تلاش نمودیم که با توسعه ابزار آریانا، که یک توالی‌یاب قدرتمند است، و در نظرگیری خواص زیستی موثر بر متیلاسیون دقت را افزایش دهیم.

## ۱-۱ تعریف مسئله

در این پژوهه ما سعی داشتیم که یک aligner برای خوانده‌های بایسولفیت‌شده بنویسیم. هدف مسئله، نگاشت یک تعداد از رشته‌های متشکل از حروف  $A, G, C, T$  بر روی یک ژنوم خاص است 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 حداًکثر و درنظر داشتن محدودیت‌های زمانی و حافظه‌ای با توجه به ابعاد مساله است.

در این پژوهه سعی ما بر آن بود که با گسترش ابزار آریانا، قابلیت هم‌ردیفی خوانده‌های بایسولفیت‌شده با ژنوم مرجع را به آن بیفزاییم. همانطور که توضیح داده شد در داده‌های بایسولفیت، سیتوزین‌هایی که متیله نباشند به تیامین تبدیل می‌شوند 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 [۲]. سعی ما در این پژوهه بر آن بود که راه حلی که ارائه می‌کنیم بر خلاف سایر aligner‌ها خواص زیستی موثر در متیلاسیون را در نظر بگیرد و با استفاده از این ویژگی بتواند دقت هم‌ردیفی را افزایش دهد.

## ۱-۲ اهداف تحقیق

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

Sequencing<sup>۱۲</sup>

Reference Genome<sup>۱۳</sup>

### ۱-۳ ساختار پایاننامه

این پایاننامه شامل پنج فصل است. فصل دوم دربرگیرندهٔ 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 اندازهٔ درخت‌ها می‌پردازیم. فصل پنجم به نتیجه‌گیری و پیشنهادهایی برای کارهای آتی خواهد پرداخت.

## ۲ فصل

### مفاهیم اولیه

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

#### ۱-۲ رشته‌ی DNA

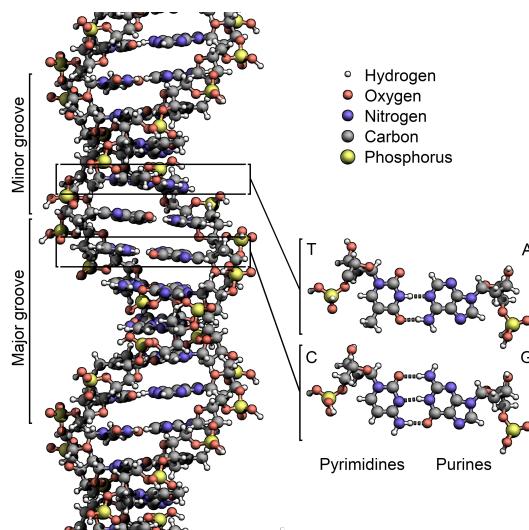
DNA مولکولی است که اطلاعات ژنتیکی مورد نیاز برای رشد و فعالیت همه ارگانیسم‌ها و برخی از ویروس‌ها را کد می‌کند. DNA نخستین بار در سال ۱۸۷۰ توسط فردريك میشر<sup>۱</sup> از هسته سلول استخراج و شناسایی گردید. DNA ساختار دو رشته‌ای دارد و این دو رشته مانند زیپ به هم متصل شده و حول یک محور مشترک پیچیده شده‌اند.

DNA یک پلیمر بسیار طویل است که از تکرار واحدهایی به نام نوکلئوتید<sup>۲</sup> به دست می‌آید. هر نوکلئوتید از یک باز نوکلئوتیدی شامل نیتروژن، تشکیل شده است که این بازها آدنین (A)، تیامین (T)، سیتوزین (C) و یا گوانین (G) هستند.

نوکلئوتیدها در یک زنجیره توسط پیوندهای کوالانسی بین قند یک نوکلئوتید و فسفات نوکلئوتید بعدی به یکدیگر متصل شده‌اند. بر اساس قوانین basepairing پیوند هیدروژنی، دو نوکلئوتید از دو رشته DNA را به هم متصل می‌کند. این پیوند به صورتی برقرار می‌گردد که ادنین و تیامین و همینطور

Friedrich Miescher<sup>۱</sup>

Nucleotide<sup>۲</sup>



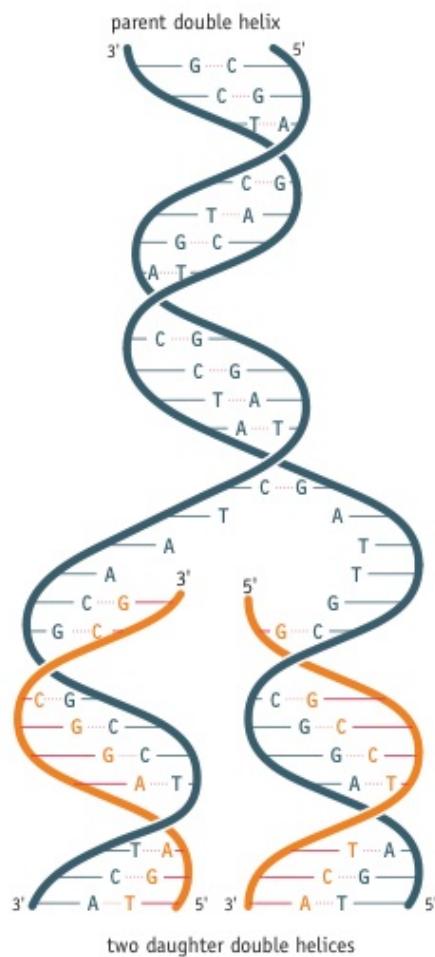
شکل ۱-۲: ساختار DNA

سیتوزین و گوانین روبروی هم قرار بگیرند.

## ۲-۲ همانندسازی DNA

برای اینکه وراثت امکان‌پذیر باشد، زن‌ها باید توانایی همانندسازی داشته باشند. زن‌ها 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.

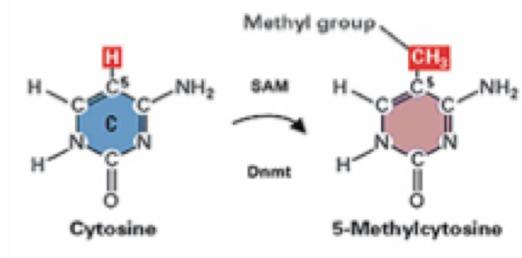
در همانندسازی DNA، دو رشته آن به کمک آنزیمی مانند زیپ از یکدیگر جدا می‌شوند و سپس از روی هر رشته، رشتهٔ جدیدی ساخته می‌شود. کلید ساخت رشتهٔ جدید در این است که روبروی هر باز، باز مکمل آن قرار می‌گیرد، به این ترتیب با استفاده از نوکلئوتیدهای آزاد که در سیتوپلاسم وجود دارند، در مقابل A، باز T و در مقابل C باز G قرار می‌گیرد و در آخر دو کپی کاملاً مشابه والد ساخته می‌شود [۳].



شکل ۲-۲: همانندسازی DNA

### 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 ۵۰۰۰ تا ۱۰۰۰۰۰ نوکلئوتید طول دارند. به مجموعه این ژن‌ها ژنوم گفته می‌شود.



شکل ۳-۲: متیلاسیون DNA

## ۴-۲ خوانده

یک مولکول DNA 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 DNA این است که مولکول بزرگ آن به قطعات کوچک شکسته شود و هر کدام از آنها جداگانه توالی‌یابی شوند. به این قطعات *fragment* یا خوانده<sup>۳</sup> می‌گویند.

## ۵-۲ متیلاسیون DNA

متیلاسیون DNA شامل اضافه شدن یک گروه متیل به انتهای کربن سیتوزین است. به سیتوزینی که گروه متیل به آن اضافه شده است سیتوزین متیله گفته می‌شود (شکل ۳-۲) [۲].

## ۶-۲ خوانده‌های بایسولفیت شده

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. CpG context :CpG context به معنی سیتوزین‌هایی می‌باشد که بلافاصله بعد از آن *G* آمده است. این سیتوزین‌ها از این جهت حائز اهمیت هستند که به دلیل خواص شیمیایی و زیستی، معمولاً متیله هستند و بسیار مورد بررسی قرار گرفته‌اند [۴].

۲. جزایر CpG: تعریف کاربردی از جزایر CpG<sup>۴</sup> به نواحی به همراه حداقل ۲۰۰ جفت باز برمی‌گردد که درصد تعداد *CG*‌ها بیشتر از ۵۰٪ باشد و نرخ مشاهده شده مورد انتظار *CpG* بیشتر از ۶۰٪ [۴]. باشد. برخلاف دستهٔ قبل، *CpG*‌های موجود در این دسته، اغلب غیرمتیله هستند [۴].

## ۷-۲ فایل Sam

فایل خروجی aligner پس از همردیفی است که در آن به ازای هر خوانده همردیف شده یک خط وجود دارد و بر اساس استراتژی aligner ممکن است خوانده‌های همردیف نشده یا خوانده‌هایی که به یک مکان یکتا همردیف نشده‌اند در آن ظاهر نشوند. در هر خط این فایل اطلاعاتی مانند نام خوانده، *ID* آن، جایگاهی که با آن همردیف شده، *flag* مشخص کننده رشته‌ای از *DNA* که با آن همردیف شده است، کروموزوم مربوط به خوانده، کیفیت خوانده و رشته سیگار<sup>۵</sup> مربوط به همردیفی آن خوانده آورده می‌شود.

## ۸-۲ رشته‌ی سیگار

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 نشان‌دهنده 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.

CpG Islands<sup>۴</sup>

CIGAR<sup>۵</sup>

## ۹-۲ هم‌ردیف‌سازی توالی

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.

## ۱۰-۲ آریانا

آریانا یک برنامه هم‌ردیف‌ساز است که از الگوریتم Burrows Wheeler برای نگاشت خوانده‌ها به ژنوم مرجع استفاده می‌کند. از نقاط برتری این هم‌ردیف‌ساز سرعت و دقیقت 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. نقطه قوت دیگر این ابزار، عدم استفاده از الگوریتم‌های Backtracking است که باعث عدم کارایی دیگر ابزارها در توالی‌یابی خوانده‌ها با عدم تطابق بالا است [۵].

## فصل ۳

# کارهای پیشین

روش‌های بسیاری برای هم‌ردیفی خوانده‌های با ایソولفیت، مانند روش‌های «wild card» و «سه حرفی<sup>۱</sup>» 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 سیتوزین‌زنوم مرچع 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 زنوم مرچع به تیامین تبدیل می‌شوند. در هر دوی این روش‌ها می‌توان نگاشت *gapped* و یا *ungapped* را بسته به برنامهٔ مورد استفاده، پیاده‌سازی کرد [۶].

### ۱-۳ الگوریتم‌های ابتدایی

برنامهٔ *CokusAlignment* از روши برای نگاشت خوانده‌ها به زنوم *Arabidopsis* که بر اساس الگوریتم‌های جستجوی درخت بنا شده است که هم از نظر محاسبات و هم از نظر حافظه بسیار ضعیف است.

Three Letter<sup>۱</sup>

با سرعت متوسط  $25 \text{ reads/sec/CPU}$  با ژنوم تقریباً کوچک اجرا می‌شود. لازم به ذکر است که می‌توان با بهینه‌سازی برای پروژه‌های مختلف این سرعت را بهبود بخشد اما در بسیاری از پروژه‌ها چنین کاری امکان‌پذیر نیست. از نظر عملی به دلیل کمبود سرعت و عملکرد، این روش قابل استفاده نیست [۷].

امروزه با توجه به الگوریتم‌های جدیدی که برای هم‌ردیفی پیاده‌سازی شده‌اند، ابزارهای هم‌ردیفی خوانده‌های بایسولفیت نیز متناسب با آن‌ها پیشرفت قابل توجهی داشته‌اند. از جمله این ابزارها ما به معرفی اجمالی BS-Seeker2 و BSmap، Bismark می‌پردازیم:

### ۲-۳ ابزار Bismark

هدف ابزار Bismark، یافتن یک تطابق منحصر به فرد با چهار بار اجرای پردازش‌ها به صورت همزمان است. در ابتدا خوانده‌های بایسولفیت با تغییراتی از نوع  $C$  به  $T$  (سیتوزین به تیامین) و از نوع  $G$  به  $A$  (گوانین به آدنین) تبدیل شده است (معادل تغییرات سیتوزین به تیامین در رشتہ معکوس). سپس هر کدام از آن‌ها به صورت معادل از نوع‌های پیش‌تغییریافته از ژنوم مرجع نگاشت می‌شوند که این عمل با استفاده از نرم‌افزار نگاشت Bowtie و به صورت چهار نمونه موازی صورت می‌گیرد. این نگاشتها، را قادر می‌سازد تا به صورت یکتا، رشتۀ اصلی بایسولفیت را مشخص نماید [۸].

### ۳-۳ ابزار BSmap

از یک فرضیه اصلی که تمام جایگاه‌های سیتوزین که تغییرات نامتقارن سیتوزین به تیامین در آنها رخ می‌دهند، شناخته شده هستند، برای راهنمایی در هم‌ردیفی خوانده‌های بایسولفیت شده استفاده می‌کند. این ابزار، تیامین‌های موجود در خوانده‌های بایسولفیت شده را به عنوان سیتوزین در نظر می‌گیرد (برعکس تغییرات بایسولفیت). که این تغییر را فقط در جایگاه‌های سیتوزین ژنوم اصلی انجام می‌دهد در حالیکه کل تیامین‌های دیگر را در خوانده‌های بایسولفیت شده بدون تغییر نگه می‌دارد. بعد از انجام این تغییرات، خوانده‌های بایسولفیت شده را به طور مستقیم بر روی ژنوم نگاشت می‌نماید.

علاوه بر موارد ذکر شده، BSMAP براساس الگوریتم کاراتر Hash table seeding کار می‌کند به این صورت که ژنوم مرجع را برای تمام  $mer - k$ ‌های ممکن شاخص‌گذاری می‌نماید که هر کدام از آن‌ها

*seed* خوانده می‌شود. برای یافتن یک نگاشت، تنها جایگاه‌هایی که به طور کامل با بخشی از خوانده منطبق شده‌اند جست و جو می‌شوند. با جست و جو در جدول *seed*‌ها، اکثر جایگاه‌های نگاشتشده دور ریخته می‌شوند و کارایی جست و جو به صورت قابل ملاحظه‌ای افزایش می‌یابد [۴].

### ۴-۲ ابزار BS – Seeker

یک ورژن به روز شده *BS – Seeker* است که نخستین aligner ای بوده است که بر اساس روش سه‌حرفی نوشته شده است و از *Bowtie* استفاده می‌کند. این ابزار هم‌ردیفی، خوانده‌های بایسولفیت را پشتیبانی می‌کند و علاوه بر *indexing* ژنوم و هم‌ردیفی خوانده‌ها، سطح متیلاسیون سیتوزین‌ها را نیز می‌تواند محاسبه کند. این ابزار می‌تواند انواع مختلف داده‌ها از جمله *WGBS/RRBS*، هم‌ردیف کند. داده‌های *RRBS* را با دقیق و کارایی بالا با *indexing* تنها نواحی لازم از ژنوم انجام می‌دهد [۹].

### ۵-۳ کاستی‌های روش‌های پیشین

به طور کلی در روش‌هایی که عدم تطابق سیتوزین به تیامین را مجاز می‌دانند، یکی از بزرگ‌ترین کاستی‌ها آن است که تعداد نگاشت‌های سیتوزین/تیامین که در یک خوانده می‌تواند شناسایی شود محدود به تعداد عدم تطابق‌های مجاز در ابزار مورد استفاده است. این تعداد می‌تواند توسط عدم تطابق‌های واقعی (*SNP*‌ها) بیشتر کاسته شود و نتیجه را بیش از قبل محدود کند.

کاستی دیگر اغلب ابزارهای موجود، عدم توجه به نتایج زیستی در الگوریتم‌های پیاده‌سازی شده است که سبب می‌شود در شرایطی که خوانده‌ها به صورت فشرده از مناطق حائز اهمیت همچون جزایر *CpG* انتخاب می‌شوند، نتایج قابل قبولی ارائه ندهند.

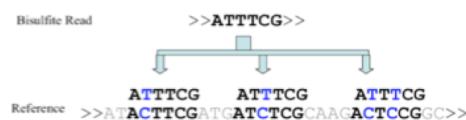
## فصل ۴

# راه حل

### ۱-۱ الگوریتم

روشی که برای هم‌ردیفی خوانده‌های بایسولفیت در aligner های امروزی مانند Brat و Bismark و BS-Seeker<sup>2</sup> رایج است تبدیل تمامی سیتوزین‌های ژنوم مرجع و تمامی سیتوزین‌های خوانده به تیامین و پس از آن هم‌ردیف کردن خوانده‌ها با ژنوم مرجع 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 متیله شده پیدا ندارند و در این ابزارها 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 (شکل ۱-۴):

- 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



شکل ۱-۴: حالات قابل قبول نگاشت C به T

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 نمی‌تواند قابل قبول 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 تمیز 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 مثبت و منفی که در واکنش با بایسولفیت تغییر یافته‌اند دیگر مکمل یکدیگر نیستند، زیرا تغییرات بایسولفیت فقط در سیتوزین‌های غیرمتیله رخ می‌دهد. 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.

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 از قاعده فوق مستثنی هستند و درصد متیلاسیون در این نواحی بسیار پایین است [۴].

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 CpG context 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 CpG context و جزایر CpG هستند را به همراه سیتوزین‌های بیرون CpG context 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 نسل جدید خطأ وجود دارد و ممکن است که در خوانده‌ای تیامین به اشتباه سیتوزین گرفته 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 CpG context 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.

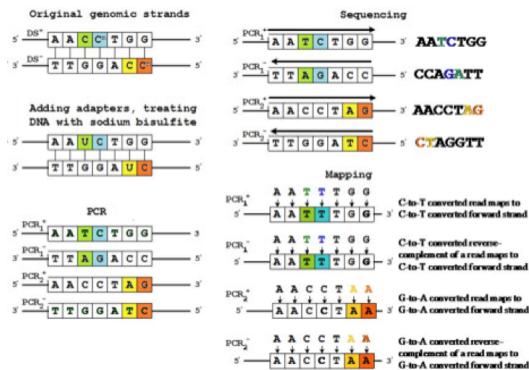
علاوه بر این دو ژنوم نیاز این دیده می‌شود که برای اینکه خوانده‌های بیشتری را بتوانیم هم‌ردیف کنیم، خوانده‌ها با ژنوم اصلی نیز هم‌ردیف شوند. دلیل آن هم این است که ممکن است نواحی‌ای وجود داشته باشند که سیتوزین‌ها علی‌رغم اینکه CpG نیستند کاملاً متیله باشند و ظاهر این نواحی مشابه ژنوم اصلی باشد. دلیل بعدی این موضوع است که عملیات بایسولفیت روی همه خوانده‌ها کامل عمل نمی‌کند و ممکن است یک سیتوزین غیرمتیله 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 CpG‌های درون جزیره را مانند دیگر CpG‌ها متیله در نظر بگیریم. به این دلیل که متیله نبودن CpG‌ها 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 دریافت نکردیم (جدول ۱-۴) و وجود آن را بی‌مورد احساس کرده و آن را از فرآیند حذف نمودیم.

این نکته قابل توجه است که در aligner‌های 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

خوانده‌ها	ژنوم اصلی	ژنوم شماره ۱	ژنوم شماره ۲	ژنوم شماره ۳
Mouse Neural Progenitor data	۱۳.۵ درصد	۴۳.۶ درصد	۱.۵ درصد	۰.۰۱ درصد

جدول ۱-۴: میزان نگاشت به هر ژنوم در رشتة مثبت



شکل ۲-۴: تبدیل PCR

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 زمینه هم‌ردیفی خوانده‌های بایسولفیت شده، 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 ۲-۴ پیداست، در این نوع، خوانده‌ها هم از رشتة مثبت و هم از رشتة منفی جمع‌آوری می‌گردد و در دستگاه PCR تکثیر می‌شوند. سیتوزین‌های غیرمتیله در این خوانده‌ها، پس از بایسولفیت شدن به تیامین تبدیل می‌شوند. به همین دلیل، اگر خوانده‌ای از رشتة منفی برداشته شده باشد، پس از بایسولفیت شدن معادل ان خواهد بود که در معکوس این خوانده گوانین‌های غیرمتیله به آدنین تبدیل شده باشند. بدیهی است که این مورد را نمی‌توان با ژنوم‌های پیشتر مطرح شده پوشش داد. به همین دلیل ژنوم‌هایی ساخته می‌شود که در آنها در رشتة 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 جزیره CpG را مد نظر قرار می‌دهیم.

## ۲-۴ انتخاب بهترین هم‌ردیفی

به منظور انتخاب بهترین هم‌ردیفی برای هر خوانده از میان هم‌ردیفی‌های مختلفی که می‌تواند با ژنوم‌های تولید شده داشته باشد، به محاسبه جریمه 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 جریمه از میان سه جریمه کم، متوسط و زیاد که مقادیر آنها ورودی برنامه هستند در نظر می‌گیریم. جریمه یک خوانده در حال حاضر جمع جریمه‌های بازهای آن است ولی برنامه نوشته شده به این صورت است که قابلیت تغییر تابع محاسبه جریمه به سادگی وجود دارد و می‌توان مدل آماری مناسب برای این کار را بدست آورد و با کمترین تغییرات ممکن آن را به برنامه افزود.

## 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. 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 متیلاسیون، خطوط SNP، خوانده‌های مصنوعی تولید کند و همچنین درصد متیلاسیون 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 متیلاسیون DNA برای کاوش هزینه، به صورت RRBS انجام می‌شود. به این صورت که درصد بسیاری زیادی از خوانده‌ها از نواحی جزایر CpG که بخش کوچکی از ژنوم را تشکیل می‌دهند بوسیله افزودن آنزیم‌هایی به DNA 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 [۹].

پس از پیاده‌سازی شبیه‌ساز، با استفاده از آن 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 فایل *Sam* نهایی خروجی برنامه، 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 فایل خط به خط خوانده می‌شود و در ابتدا اگر خوانده‌های *PCR* هم در فایل موجود باشند، بر اساس *flag* مشخص کننده رشته‌ای 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 هم‌ردیفی ۱۶
- تعداد تبدیلهای *A* به *G* بیشتر از *C* به *T* و هم‌ردیفی ۰
- تعداد تبدیلهای *A* به *G* کمتر از *C* به *T* و 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 رشتۀ منفی هم‌ردیف شده است و صفر نشان‌دهنده هم‌ردیفی با رشتۀ مثبت است. 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 رو布رو هستیم که در این حالت هم می‌توان از آنها صرف نظر کرد و هم به صورت تصادفی 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 *C* های غیر متیله اضافه می‌شود. در واقع پس از پایان این فرآیند برای تمامی خطوط فایل *Sam* اطلاعاتی که ما به ازای هر سیتوزین در ژنوم مرجع خواهیم داشت عبارتند از:

- جایگاه سیتوزین نسبت به ابتدای کروموزوم
- رشته‌ای که خوانده با آن هم‌ردیف شده است
- تعداد خوانده‌هایی که این سیتوزین را پوشش داده و در آنها این  $C$ ، بوده است. (تعداد خوانده‌های مตیله)
- تعداد خوانده‌هایی که این سیتوزین را پوشش داده و در آنها این  $C$ ،  $T$  بوده است. (تعداد خوانده‌های غیرمتیله)

و با داشتن این مقادیر، درصد متیلاسیون هر سیتوزین به صورت جداگانه برای هر رشته بدست می‌آید.

## فصل ۵

### پیاده‌سازی

در اولین گام پیاده‌سازی لازم است که ژنوم‌های مورد نیاز (که در فصل را محل 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 هم‌ردیفی 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 بايسولفیت و غیر بايسولفیت تفاوت قائل شود. در حالتی که اجرای بايسولفیت مد نظر 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 فایل *Sam* است که حاصل ۵ بار هم‌ردیفی است. در گام آخر باید از میان هم‌ردیفی‌های صورت گرفته (۵ فایل) بهترین را انتخاب کنیم. بدیهی‌ترین روش برای انجام این مورد آن است که به ازای هر خواننده، تمامی فایل‌ها 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 خواننده تمامی فایل‌ها را یک بار جستجو کرد. به همین منظور در ابتدا فایل‌ها را بر اساس نام خواننده‌ها، توسط دستور *sort* سیستم‌عامل لینوکس<sup>۱</sup> مرتب می‌کنیم. سپس هر سطر از هر ۵ فایل را می‌خوانیم و از میان آن‌ها بهترین

Linux<sup>۱</sup>

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 سطrix در فایل نهایی قرار می‌دهد. به همین دلیل، ترتیب خوانده‌ها پس از مرتب‌سازی در تمامی ۵ فایل یکسان است.

## ۱-۵ تابع جریمه

تشخیص نواحی CpG context بر این اساس است که آیا بعد از هر سیتوزین در رشته مثبت گوانین آمده است یا خیر و در صورت مشاهده *CG* یک ناحیه در نظر گرفته می‌شود. مکان‌های جزایر *CpG* را در حافظه به صورت مرتب شده ذخیره کرده و برای بررسی کردن اینکه یک سیتوزین در جزیره *CpG* قرار دارد یا خیر بر روی داده ساختاری که مکان شروع و پایان جزایر را نگه داشته، جستجوی دودویی صورت می‌پذیرد.

در تابع فعلی برای محاسبه جریمه بازها، اگر باز در خوانده و ژنوم هر دو *C* و بعد از آن *G* آمده باشد و در جزیره *CpG* قرار گرفته باشد، جریمه متوسط (زیرا معمولاً سیتوزین‌های درون جزایر غیر متنیله‌اند) و در خارج جزیره جریمه کم اختصاص می‌یابد (به این دلیل که در این حالت انتظار می‌رود خوانده متنیله باشد پس سیتوزین، بعد از بایسولفیت سیتوزین باقی می‌ماند) و در خارج CpG context جریمه زیاد اختصاص می‌یابد (به این دلیل که در این حالت انتظار می‌رود خوانده متنیله نباشد).

در صورتی که باز در خوانده *T* و مقدار متناظر در ژنوم *C* باشد در حالت اول جریمه صفر و در حالت دوم جریمه کم اختصاص داده می‌شود (به این دلیل که در این حالت انتظار می‌رود خوانده متنیله باشد پس سیتوزین، بعد از بایسولفیت باید سیتوزین باقی بماند) و در خارج CpG context جریمه کم اختصاص می‌یابد (به این دلیل که در این حالت انتظار می‌رود خوانده متنیله نباشد).

برای حالات درج و حذف نیز بیشترین جریمه 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 جریمه کلی یک خوانده محاسبه می‌شود.

## ۲-۵ محاسبه درصد متیلاسیون

به دلیل ابعاد بزرگ مساله و محدودیت حافظه‌ای که وجود دارد خوانده‌های هم‌ردیف شده خط به خط از فایل ورودی گرفته می‌شوند و جایگاه سیتوزین‌ها به همراه مقادیر توضیح داده شده در یک بافر<sup>۲</sup> حلقوی به سایز طول بزرگترین خوانده، 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 نگاشت سیتوزین‌های ژنوم به بافر از یک تابع *hash* خطی ساده استفاده می‌شود.

## فصل ۶

# نتایج

در این فصل به بررسی عملکرد ابزار تهیه شده می پردازیم و این ابزار را از جهات مختلف با دیگر ابزارهای موجود مورد قیاس قرار می دهیم.

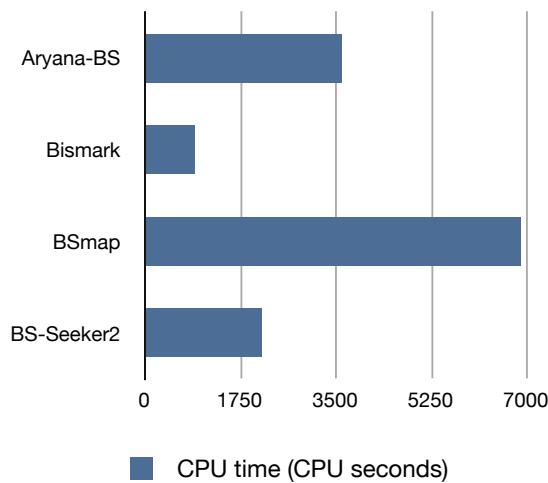
### ۱-۶ مقایسه زمان اجرا

در نمودار ۱-۶ زمان اجرای هر سه ابزار BSmap، Bismark و BS-Seeker2 برای داده *benchmark*<sup>۱</sup> Bismark که شامل یک میلیون خوانده به طول ۸۷ و با یسولفیت شده از ژنوم hg19 است، اندازه گرفته و با زمان اجرای آریانا مقایسه می شود که همانطور که از نمودار مشخص است، زمان اجرای آریانا از Bismark و BS-Seeker2 بیشتر بوده است و آن هم به دلیل این است که آریانا خوانده ها را با تعداد ژنوم بیشتری از این دو ابزار هم ردیف می کند (۵ ژنوم).

زمان اجرای BSmap نیز به علت ساخت k-mer ها از ژنوم بسیار بالا بوده است. زمان اجرای هم ردیفی، به وسیله دستور time اندازه گیری و همه aligner ها با ۸ ریسه و روی محیط یکسان با مشخصات ذکر شده اجرا شده اند.

---

<http://www.cbrc.jp/dnemulator/bab/><sup>۱</sup>



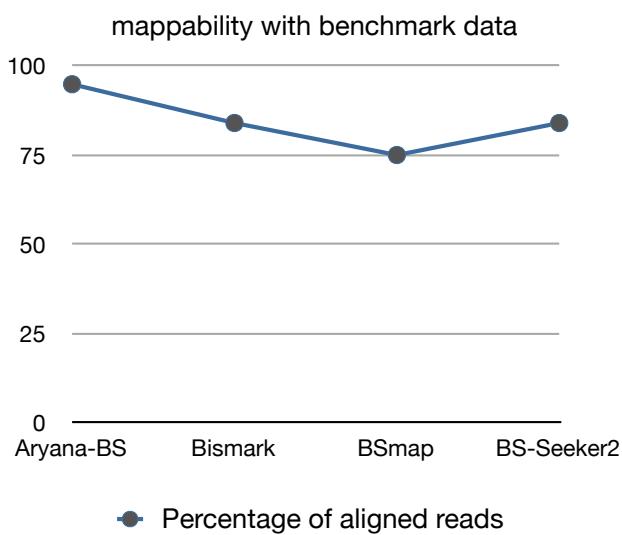
شکل ۱-۶: مقایسه زمان اجرا

## ۲-۶ مقایسه توانایی هم‌ردیف‌سازی و دقت

در نمودار ۲-۶ درصد خواندهای هم‌ردیف شده توسط هر aligner نمایش داده شده است. این درصدها برای خواندهای benchmark توضیح داده شده، به دست آمداند و همانگونه که مشخص است آریانا بیشترین تعداد خوانده را هم‌ردیف کرده است.

در نمودار ۳-۶ درصد خواندهای هم‌ردیف شده به جایگاه‌های صحیح توسط هر aligner نمایش داده شده است. این درصدها برای خواندهای benchmark توضیح داده شده و با مقایسه جایگاه نوشته شده در فایل‌های خروجی با جایگاه‌های اصلی که به همراه داده وجود داشت، بدست آمداند و برای آریانا، Bismark، BSmap و BS-Seeker2 برابر با ۸۲.۵، ۵۵.۲، ۸۱.۲ و ۷۲.۲ بوده است که آریانا و بعد از آن Bismark بیشترین دقت هم‌ردیفی را داشته‌اند.

در نمودار ۴-۶ نیز درصدهای کل خواندهای هم‌ردیف شده و خواندهای صحیح هم‌ردیف شده مشاهده می‌شود.



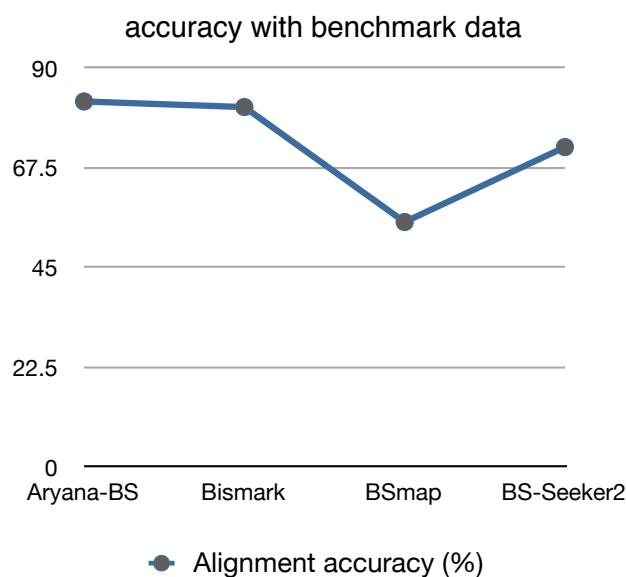
شکل ۶-۲: مقایسه توانایی هم ردیف سازی

### ۳-۶ مقایسه توانایی هم ردیف سازی خوانده های *PCR*

داده های نمودار ۵-۶، خوانده هایی هستند که توسط شبیه ساز ما، تولید شده اند. خوانده های این داده *PCR* شده و تعداد آنها ۱ میلیون و فاقد خطأ هستند. در این نمودار درصد خوانده های هم ردیف شده و دقت هم ردیفی مشاهده می شود. آریانا با اختلاف زیادی در دقت هم ردیفی پیش رو است (۹۵.۸٪) و در جایگاه بعدی BSmap با ۴۸٪ دقت قرار دارد.

### ۴-۶ بررسی دقت محاسبه درصد متیلاسیون

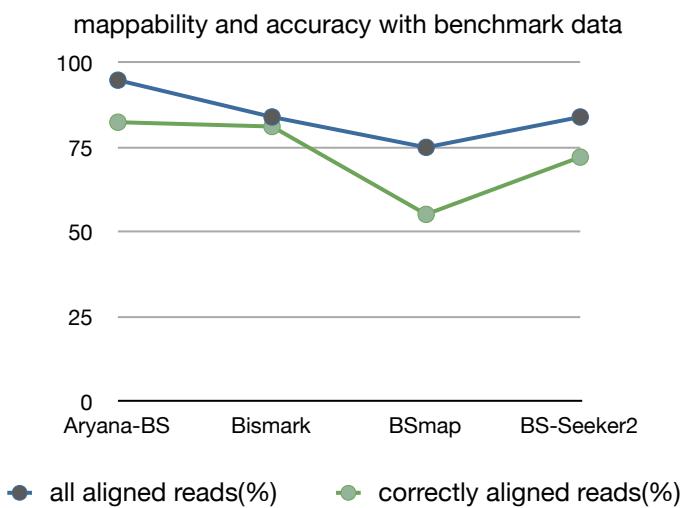
داده های نمودار ۶-۶، خوانده های شبیه سازی شده از کروموزوم ۱۰ انسان هستند. تعداد این خوانده ها ۶۰ میلیون است تا پوشش ژنوم انتخاب شده بالا باشد، همچنان ۳۰٪ از این خوانده ها از جزایر CpG داده شده اند که متیلاسیون آنها اهمیت بالایی دارند. در این نمودار محور X معادل درصد متیلاسیون خروجی از شبیه سازی و محور Y معادل درصد های محاسبه شده توسط آریانا است. همانطور که انتظار می رفت ۳ دسته کلی برای درصد متیلاسیون سیتوزین ها وجود دارد که شامل سیتوزین های خارج CpG



شکل ۳-۶: مقایسه دقت هم‌ردیف‌سازی

با ۱ درصد متیلاسیون، سیتوزین‌های داخل CpG context و بیرون از جزیرهٔ CpG با ۹۰ درصد متیلاسیون و سیتوزین‌های درون جزیره با ۳ درصد متیلاسیون هستند. همان‌گونه که انتظار می‌رفت اکثر نقاط بر روی خط با شیب ۱ قرار گرفته‌اند که این نشان می‌دهد که آریانا در اکثر موارد، درصد درستی را محاسبه کرده است. تعدادی از نقاط نیز در مکان‌های نادرستی قرار گرفته‌اند که این می‌تواند ناشی از نرخ خطای ۲ درصد و همچنین ۲ میلیون SNP در خوانده‌ها باشد.

داده‌های نمودار ۷-۶ نیز خوانده‌های شبیه‌سازی شده قبلی هستند. در این نمودار فراوانی اختلاف درصد متیلاسیون محاسبه شده و درصد متیلاسیون واقعی نشان داده شده است و شکل نمودار و قله با ارتفاع بسیار بلند در صفر بیانگر این است که درصدهای متیلاسیون محاسبه شده بسیار نزدیک به درصدهای واقعی هستند. همانطور که مشهود است با دور شدن از مبدأ و افزایش خطا کاهش چشمگیری در فراوانی درصد اختلاف‌ها رخ می‌دهد و نسبت تعداد آنها به کل خوانده‌ها به صفر میل می‌کند.

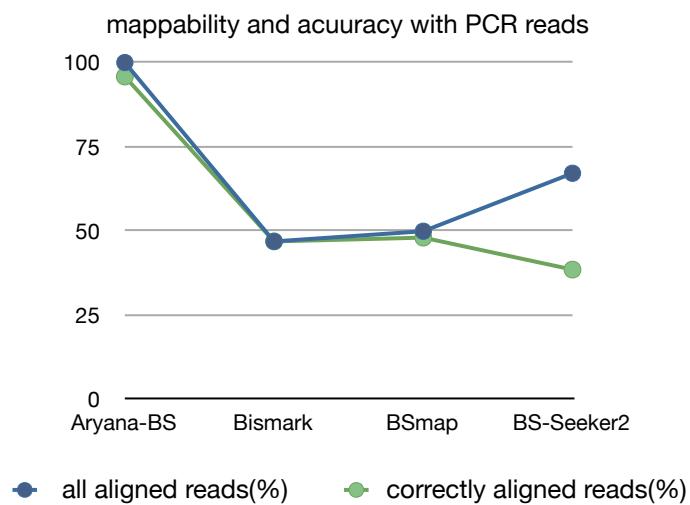


شکل ۴-۶: مقایسه دقت و توانایی هم‌ردیف‌سازی

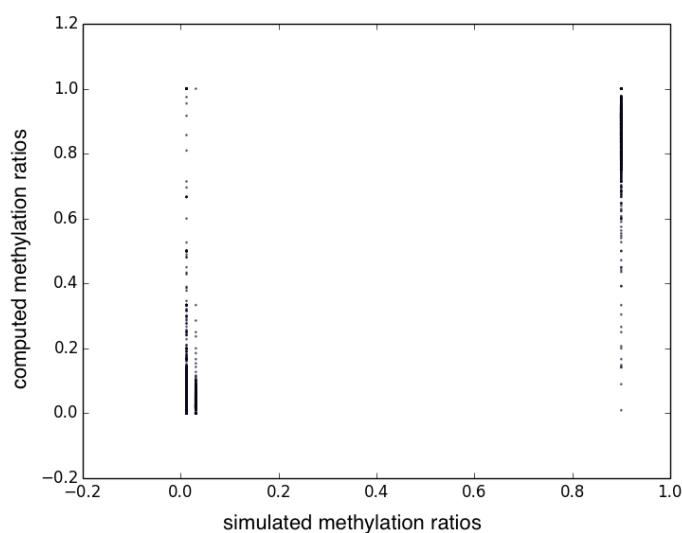
## ۵-۶ کارهای آینده

ابزار آریانا این قابلیت را دارد که از ابعاد مختلف گسترش داده شود. تلاش ما بر این است که تا جای ممکن به قابلیتها و امکانات و همچنین دقت هم‌ردیفی خوانده‌های بایسولوژیت شده بیفزاییم و در ادامه ابعاد دیگری از مسائل مربوط به توالی‌یابی را با ابزار آریانا پشتیبانی کنیم.

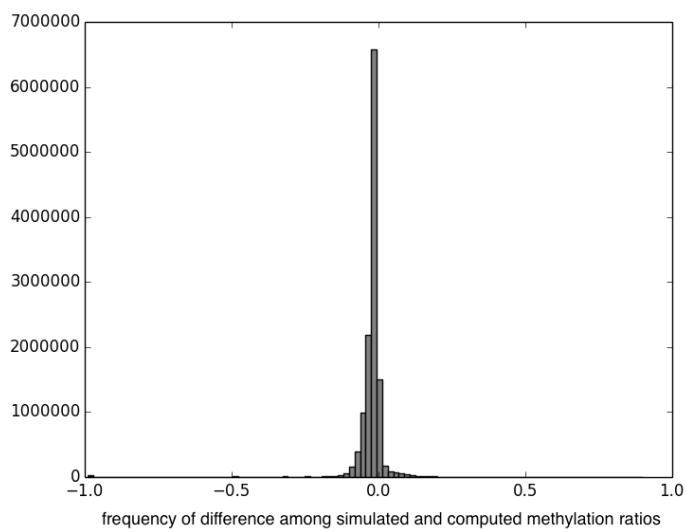
از دیگر مواردی که تصمیم داریم در آینده به این ابزار بیفزاییم، استفاده از مدل‌های آماری پیچیده و کاراتر در محاسبه جریمه و انتخاب بهترین هم‌ردیفی است که می‌تواند تأثیر بهسزایی در افزایش دقت داشته باشد.



شکل ۶-۵: مقایسه دقت در خواندهای PCR



شکل ۶-۶: درصد متیلاسیون واقعی و به دست آمده



شكل ٦-٧: فراوانی اختلاف درصد متیلاسیون

# كتاب نامه

- [1] Martin C Frith, Ryota Mori, and Kiyoshi Asai. A mostly traditional approach improves alignment of bisulfite-converted dna. *Nucleic acids research*, page gks275, 2012.
- [2] Felix Krueger, Benjamin Kreck, Andre Franke, and Simon R Andrews. Dna methylome analysis using short bisulfite sequencing data. *Nature methods*, 9(2):145–151, 2012.
- [3] Terry Brown. *Introduction to genetics: a molecular approach*. Garland Science, 2012.
- [4] Alexander Meissner, Tarjei S Mikkelsen, Hongcang Gu, Marius Wernig, Jacob Hanna, Andrey Sivachenko, Xiaolan Zhang, Bradley E Bernstein, Chad Nusbaum, David B Jaffe, et al. Genome-scale dna methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770, 2008.
- [5] Milad Gholami, Aryan Arbab, Ali Sharifi-Zarchi, Hamidreza Chitsaz, and Mehdi Sadeghi. Aryana: Aligning reads by yet another approach. *BMC bioinformatics*, 15(Suppl 9):S12, 2014.
- [6] Govindarajan Kunde-Ramamoorthy, Cristian Coarfa, Eleonora Laritsky, Noah J Kessler, R Alan Harris, Mingchu Xu, Rui Chen, Lanlan Shen, Aleksandar Milosavljevic, and Robert A Waterland. Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. *Nucleic acids research*, page gkt1325, 2014.
- [7] Yuanxin Xi and Wei Li. Bsmap: whole genome bisulfite sequence mapping program. *BMC bioinformatics*, 10(1):232, 2009.
- [8] Felix Krueger and Simon R Andrews. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, 27(11):1571–1572, 2011.
- [9] Weilong Guo, Petko Fiziev, Weihong Yan, Shawn Cokus, Xueguang Sun, Michael Q Zhang, Pao-Yang Chen, and Matteo Pellegrini. Bs-seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC genomics*, 14(1):774, 2013.



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