بسم الله الرحمن الرحیم

جزوه ی درس مبانی بیوانفورماتیک.

استاد : دکتر مطهری

سید امیر حسین صابری

جلسه ی سوم از بحث denovo sequencing :

۱. مقدمه:

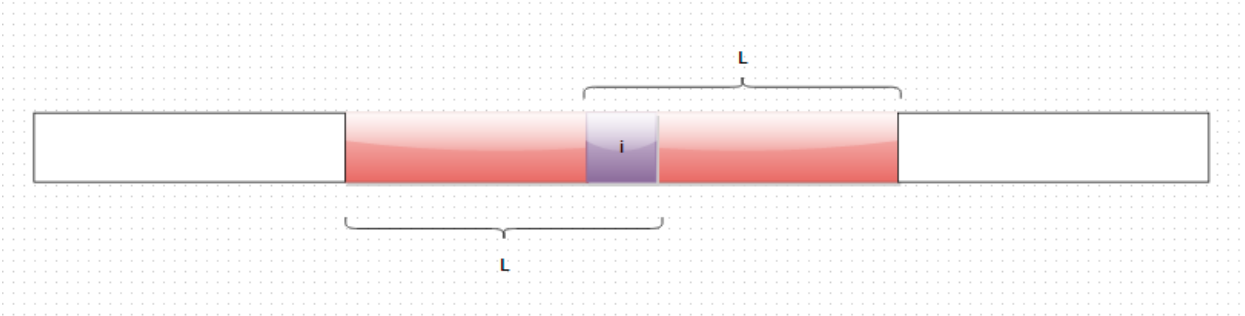
اگر ما از ژنوم بی نهایت read برداریم که نویزی هم نیستند در این صورت می توانیم ادعا کنیم که L-spectrum را داریم.ولی حتی اگر چنین شرایطی هم برقرار باشد،اگر طول read از L_{crit} (که اگر خانه های ژنوم به صورت i.i.d باشند داریم $L_{crit} = \frac{2*\log(G)}{H2(p)}$) کوچکتر باشد باز هم نمی توان ژنوم را یافت.(به دلیل این که احتمال دارد در ژنوم تکرار سه تایی یا interleaved داشته باشیم.

حال قصد داریم تعداد read هایی را بیابیم که بتوانیم ادعا کنیم می شود ژنوم را بازسازی نمود.

۲. بدست آوردن آستانه ی تعداد read :

E = {رشته ی ژنوم پوشیده نشده باشد} = $\cup E_i | E_i = \{$ باز iام پوشش ندارد $\}$

$P(E) \leq \sum_{i=1}^{G-L+1} P(E_i) \leq G*P(E_i)$



با توجه به شکل بالا در میابیم که برای این که یک خانه پوشانده نشود باید L تا از read ها مورد استفاده قرار نگیرند.

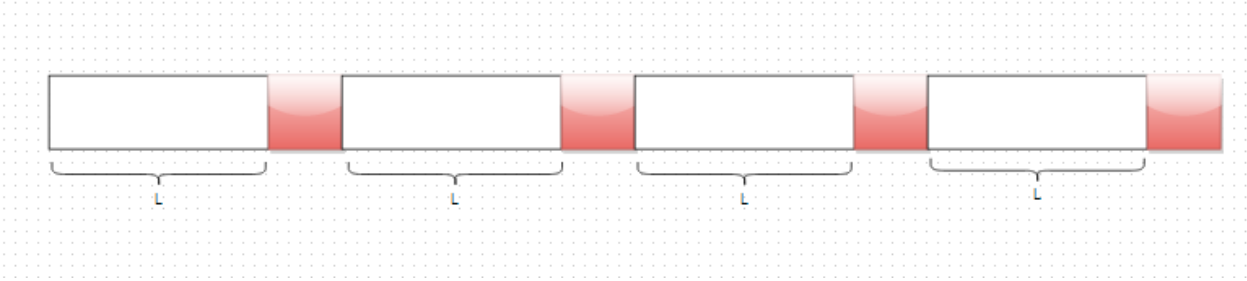
بنابراین داریم:

$G*P(E_i) = G*(1-\frac{L}{G})^N \leq Ge^{-\frac{NL}{G}} \leq e^{\ln G - \frac{NL}{G}}$

برای این که این عبارت به سمت صفر رود باید داشته باشیم:

$\frac{NL}{G} \gg \ln G \rightarrow N \gg \frac{G \ln G}{L} = N_{coverage}$

حال می خواهیم اثبات کنیم که اگر تعداد read ها از این مقدار کم تر شود پوشش نداریم.



مطابق شکل به فاصله ی L بیس ، خانه هایی را تعیین می کنیم.می خواهیم ببینیم که به ازای چه تعداد read همین (تعداد $\frac{G}{L}$ از) بیس هایی که مشخص کرده ایم بدون پوشش باقی می مانند.توجه شود که coverage این خانه ها از هم مستقل می باشد.

که این مسئله همان مسئله ی معروف coupon collector می باشد.و جواب آن می شود:

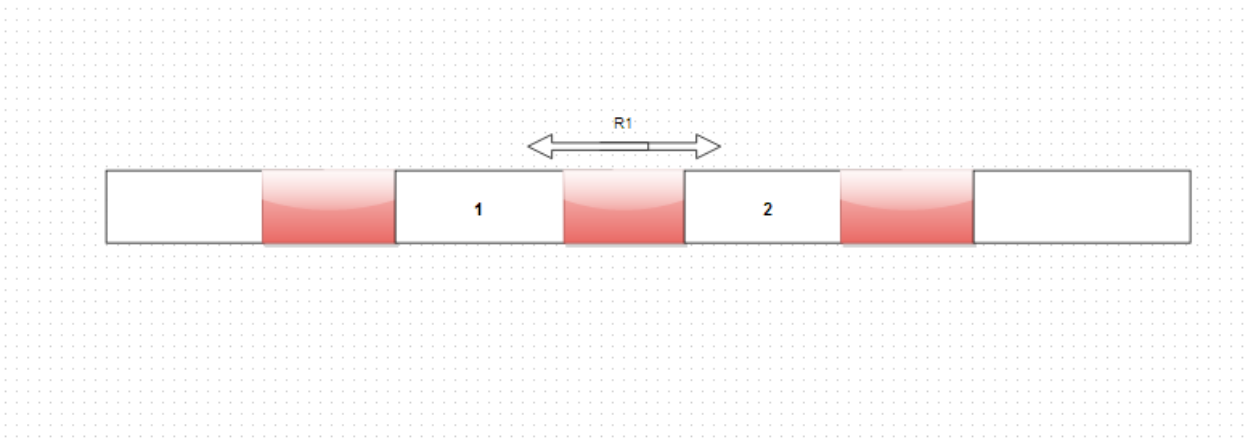
$$N < \frac{G}{L}\log\frac{G}{L}$$

از آن جا که L نسبت به G خیلی کوچک است.سمت راست معادله می شود $\frac{G\ln G}{L}$ که همان $N_{coverage}$ می باشد که در بالا بدست آوردیم.

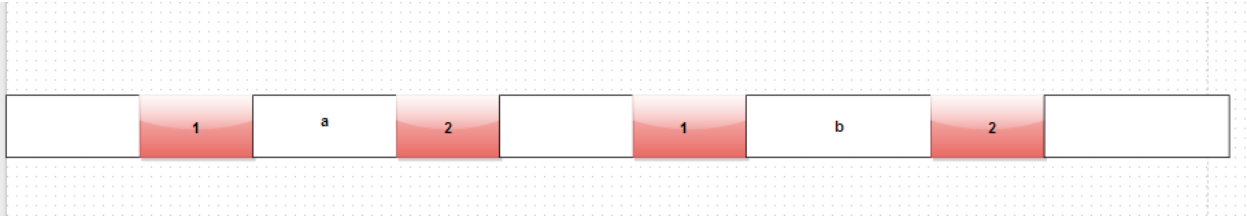
بنابراین اگر تعداد read ها از $N_{coverage}$ کمتر باشد نمی توان همه ی بیس ها را پوشش داد.

قضیه:

اگر ما یک تکرار سه تایی یا یک interleaved داشته باشیم که هیچ readای آن را bridge نکرده باشد آن گاه نمی توان آن را بازسازی کرد.



اگر قسمت شماره ی ۱ و شماره ی ۲ را در شکل بالا جا به جا کنیم تغییری در L-spectrum ایجاد نمی شود مگر اینکه R1، read وجود داشته باشد.در این صورت جا به جایی دو قسمت ذکر شده L-spectrum را تغییر خواهد داد.



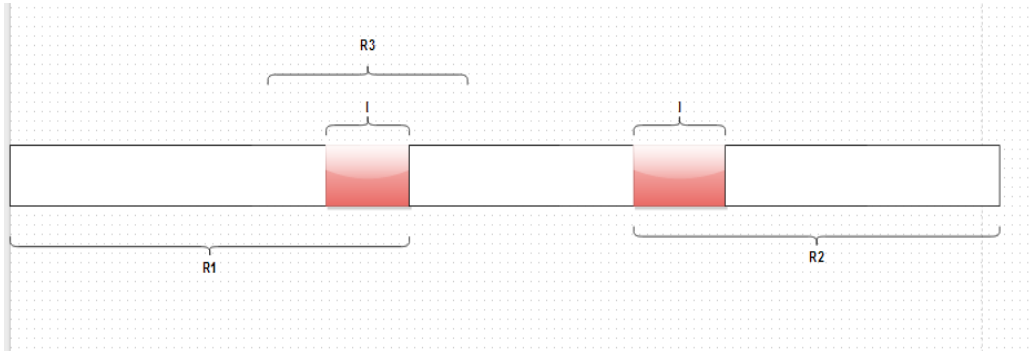
برای تکرار interleaved هم،همانطور که دیده می شود مثل حالت سه تایی است.

۳. الگوریتم greedy :

دو read ای را که با هم بیشترین overlap را دارند به هم می چشبانیم و این کار را آنقدر ادامه می دهیم تا تنها یک read باقی بماند.

قضیه:

الگوریتم greedy رشته را بازسازی می کند اگر هر تکراری توسط حداقل یک read پوشانده شده باشد.(اصلاح : هر بیس داخل(ابتدا یا انتها نه) یک read باشد.)



همانطور که در شکل بالا می بینیم اگر repeat به طول L که با رنگ قرمز نشان داده شده توشط یک read ، bridge نشود آنگاه دو read، R1 و R2 طبق الگوریتم greedy به هم وصل می شوند که غلط است.ولی اگر برای مثال read، R3 هم وجود داشته باشد در این صورت الگوریتم R1 و R3 را به هم وصل می کند که کاملا درست است. بنا براین الگوریتم greedy به ازای $l > L^*$ و به ازای $N > N^*$ به درستی کار می کند.(که N^* تعداد read هایی می باشد که علاوه بر پوشش همه ی repeatها bridge می شوند و L^* طولی می باشد که به ازای آن علاوه بر این که triple و interleaved نداریم repeat هم نداریم.)

اگر بیس های ژنوم به صورت i.i.d باشند در این صورت داریم:

$N^* = N_{coverage}$

$L^* = L_{crit}$

۴. محاسبه ی احتمال خطا:

احتمال خطا:احتمال اینکه یک repeat به طول l داشته باشیم و bridge هم نشده باشد.

احتمال repeat به طول l = $2^{-lH_2(P)}$

اجتمال عدم پل زدن = $(1-2\frac{L-l}{G})^N$

احتمال خطا = $\sum_{l=0}^{L} G^2\,(1-2\frac{L-l}{G})^N 2^{-lH_2(P)} \le \sum_{l=0}^{L} G^2\, e^{-2\frac{N}{G}(L-l)} 2^{-lH_2(P)}$

$l=0 \;\rightarrow\; G^2 e^{-2\frac{N}{G}(L-l)} \downarrow 0 \;\rightarrow N \gg \frac{G}{L}\ln G \quad (N_{coverage})$

$l=L \;\rightarrow G^2 e^{-lH_2(p)} \downarrow 0 \;\rightarrow L \gg \frac{2\log G}{H_2(p)} \quad (L_{crit})$

۵. الگوریتم ساده ی k-mer:

در این الگوریتم read را به تکه های k بیسی می شکنیم و به نوعی k-spectrum تشکیل می دهیم.
شروط:

a. $K > L_{crit}$
b. هر k تایی با حد اقل یک read پوشانده شود.

(شرط سنگینی می باشد،اصلاح:دو read مجاور حد اقل k ، overlap دارند.

این الگوریتم به L بلند یا N خیلی زیاد نیاز دارد.