به نام خدا

تعریف مسئله:

هدف مقاله ارایه روشی برای مقایسه‌ی دو ژن با طول زیاد و همچنین ارایه‌ی نمودار dot-plot و معیاری برای سنجش میزان شباهت در زمان و حافظه‌ی ثابت (وابسته به دقت مقایسه). از جمله ویژگی‌ها این الگوریتم می‌توان به موارد زیر اشاره کرد:

- جستجوی شباهت‌های دو ژنوم:
  - مقایسه‌ی دو ژنوم خیلی طولانی
  - Synteny Map برای ۱۲ گونه از پستانداران
- بررسی رویدادهای تکاملی در چند گونه و پیدا کردن تیکه‌های مشابه در ژنوم‌ها و پی بردن به رویدادهای تکاملی بین گونه‌ها
- مصرف پایین از قدرت پردازشی و حافظه به نسبت الگوریتم‌های دیگر

تعاریف اولیه:

شامل:

- مجموعه‌ی حروف: S = {A, C, G, T}
- هر رشته در این زبان عبارت است از: [ACGT]+
- اگر s یک رشته به طول n در این زبان باشد، مجموعه‌ی همه‌ی زیر رشته‌های به طول k را C می‌نامیم. (مجموعه تمام k-merها)
- $H_{a,b}$ مجموعه‌ی تمام زوج k-merهای مساوی، در حاصل ضرب دو مجموعه‌ی C_a و C_b که به آن مجموعه hits گفته می‌شود.

همه‌ی زوج k-merهای حاصل ضرب دو مجموعه‌ی C_a و C_b از ۳ توزیع زیر می‌آیند:

- $W_{a,b}$: توزیع این است که k-merای از رشته‌ی اول به طور تصادفی دقیقا در رشته‌ی دوم نیز آمده باشد که از توزیع binomial پیروی می‌کند.
- $R_{a,b}$: توزیع برخوردهای که شامل دو k-merای باشد که بیش از یکبار در ژنوم تکرار شده باشد.

- $T_{a,b}$: توزیع k-merهای محافظت شده(Conserved) که شامل همه‌ی زوج k-merها به جز k-merهای دو توزیع قبل است که آنها نویز هستند.

ساخت شاخص غیر دقیق برای k-merها:

برای اینکار از تابع زیر استفاده شده است:

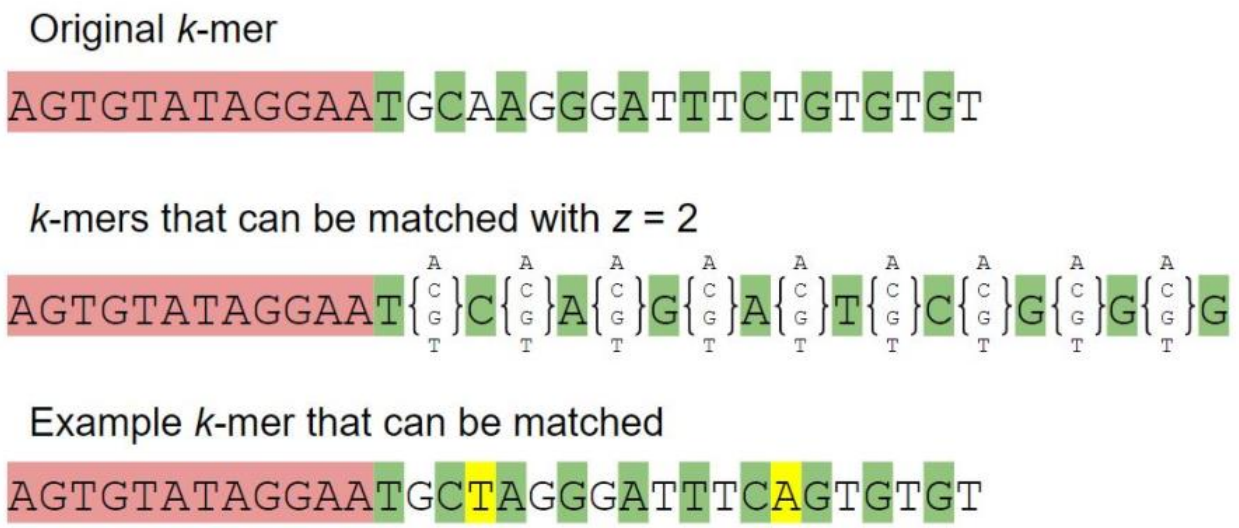
$$f(w_i^k) = |S|^0 * v(w_{i,1}^k) + |S|^1 * v(w_{i,2}^k) + \dots + |S|^{j-1} * v(w_{i,j}^k) + \dots + |S|^{k-1} * v(w_{i,k}^k)$$

$$f(w_i^k) = \sum_{j=1}^{k} |S|^{j-1} * v\left(w_{i,j}^k\right)$$

تابع بالا شاخص دقیقی برای یک رشته است. بر روی این شاخص می‌توان یک سری تابع شاخص گذاری غیر دقیق به صورت زیر تعریف کرد:

$$f(w_{i,j}^k)^z = \sum_{j=1}^{k} (|S|^{j-1} * v(w_{i,j}^k) * belongs(j, z))$$

که belongs(j, z) به معنی این است که j مضربی از z باشد. با انتخاب z بزرگتر از یک، تابع بالا یکسری از حروف را در شاخص گذاری تاثیر نمی‌دهد در نتیجه تعداد برخوردها بیشتر می‌شود زیرا بعضی از k-merهایی که برابر نیستند، شاخص یکسانی می‌گیرند.



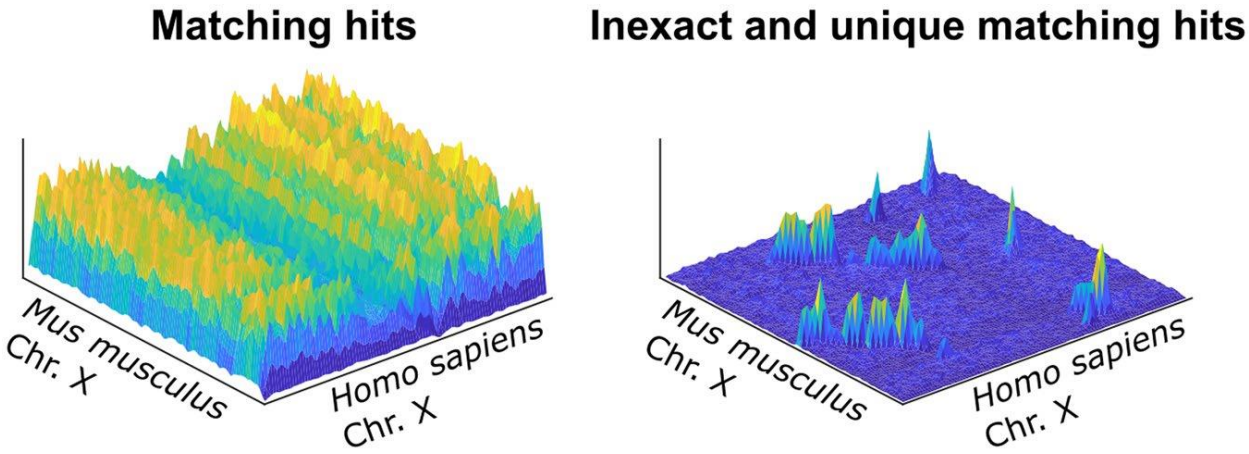
همچنین شاخص گذاری باعث بالا رفتن سرعت پیدا کردن k-merهای (تقریبا) یکسان می‌شود.

فیلترکردن ابتکاری برخوردها:

از آنجایی که احتمال برخورد دو k-mer به صورت رندوم در ژنوم‌های طولانی و با k مناسب، بسیار پایین است، در نتیجه می‌توان این برخوردها را حذف کرد. این برخوردها شامل توزیع $W_{a,b}$ هستند.

همچنین برخوردهایی که در توزیع $R_{a,b}$ هستند نیز از نظر بیولوژی احتمال کمی دارند. در نتیجه این برخوردها را نیز از مجموعه‌ی برخوردها حذف می‌کنیم.

با حذف برخوردهای گفته شده، تنها برخوردهایی که متعلق به محل‌های محافظت شده‌اند باقی می‌ماند. البته ممکن است در اینها نیز مقداری نویز باشد که با کوچک کردن(down-sampling) ابعاد سعی در از بین بردن نویزها می‌کند.



تابع امتیاز دهی:

همچنین معیاری برای میزان شباهت دو ژنوم ارایه شده که با استفاده از ماتریس برخوردهای محافظت شده که در بخش قبل توضیح داده شد بدست می‌آید. این معیار بیشتر به حفظ محل‌های محافظت شده(Conserved Region) امتیاز می‌دهد. در این معیار که بین صفر تا یک است، هرچه به صفر نزدیکتر باشد به معنی آن است که دو ژنوم

$$d_{raw} = \sum_{i}^{l-1} taxicab(\max(H_m(i), H_m(i+1)))$$

$$d = \frac{d_{raw}}{l^2}$$

کاملا شبیه هم هستند و قطر اصلی در dot-plot کاملا حفظ شده است و هرچه به یک نزدیکتر باشد به معنی آن است که دو ژنوم شباهتی به هم ندارند.

مراحل اجرای الگوریتم:

۱. شاخص گذاری غیر دقیق k-merهای یکتا در رشته‌ی مرجع
۲. پیدا کردن برخوردها بین رشته‌ی مرجع و رشته‌های query با استفاده از همان شاخص گذاری
۳. حذف برخوردهای غیر یکتا
۴. کاهش ابعاد(down-sampling) ماتریس برخوردها
۵. نگه داشتن قله‌ها و حذف باقی برخوردها از ماتریس برخوردها
۶. محاسبه‌ی فاصله‌ی بین رشته‌ها با استفاده از تابع امتیاز دهی گفته شده
۷. پیدا کردن بازآرایی‌های ژنوم‌ها در ابعاد بزرگ(LSGRs) با بررسی خط‌های بوجود آمده در ماتریس برخوردها