دانشکده مهندسی کامپیوتر
دانشگاه صنعتی شریف

ژنومیک محاسباتی

مطهری و فروغمند
پاییز ۱۴۰۰

هم‌ردیفی چندگانه

جلسه هشتم

نگارنده: ریحانه صادقی

۱ مروری بر مباحث گذشته

در هم‌ترازی دوگانه که در جلسات قبل به آن اشاره شد، دو رشته داریم که هم‌ترازی بین آن‌ها را بر اساس یک ماتریس δ که یک ماتریس مشابهت حروف است، به دست می‌آید.

در خروجی آن نیز یک جدول داریم که دارای دو سطر(رشته) است که سطر اول آن، همان رشته S و سطر دوم آن، رشته T است که به هرکدام تعدادی «ـ» به عنوان فضای خالی (Gap) اضافه شده است. نتیجه مسئله هم‌ترازی دو رشته‌ای است که بالاترین امتیاز را بر اساس ماتریس δ به دست بیاورد.

۲ چرا هم‌ترازی

تغییراتی در DNA رخ می‌دهد و در هم‌ترازی می‌خواهیم آن‌ها را تشخیص دهیم. در هم‌ترازی به دنبال اطلاعات مهم می‌گردیم. به دنبال شباهت‌های مهم که موجب عملکرد یکسان می‌شوند. مثلا تفاوت‌هایی بین برخی ژن‌ها وجود دارد در حالی که آن ژن‌ها عملکرد یکسانی دارند که ما به دنبال این شباهت های مهم که باعث عملکرد یکسان می‌شوند، می‌گردیم.

دقت شود بررسی همه حالت ها بسیار سخت است بنابراین در طبیعت به دنبال شباهت‌ها بین اجزایی که متفاوت هستند اما عملکرد یکسان دارند می‌گردیم تا بتوانیم آن پارامترهای مهم که عملکرد را حفظ می‌کنند، پیدا کنیم. از این رو به هم‌ترازی چندگانه نیاز پیدا می‌کنیم.

S_{11}	-	S_{12}	S_{13}	S_{14}	-	-	-	S_{15}	-
-	S_{21}	S_{22}	S_{23}	-	-	S_{24}	S_{25}	S_{26}	S_{27}
S_{31}	S_{32}	-	S_{33}	-	-	S_{34}	S_{35}	-	S_{36}

شکل ۱: نمونه‌ی خروجی هم‌ترازی چندگانه

۳ هم‌ترازی چندگانه

در هم‌ترازی چندگانه چند رشته S_i و ماتریس مشابهت حروف به عنوان ورودی مسئله داده می‌شود و در خروجی، مشابه هم‌ترازی دو گانه، یک جدول با i سطر داریم که هر سطر i، همان رشته S_i است که به هرکدام تعدادی «ـ» به عنوان فضای خالی (Gap) اضافه شده است. نمونه‌ای از این خروجی در شکل ۱ مشاهده می‌شود. نتیجه مسئله هم‌ترازی رشته‌هایی است که بالاترین امتیاز را بر اساس ماتریس δ به دست بیاورد. برای به دست آوردن این امتیاز، به ازای هر ستون، مجموع امتیاز دو به دوی سطرها را به دست آورده و مجموع امتیاز همه ستون‌ها را به عنوان امتیاز نهایی محاسبه می‌کنیم. دقت شود که در محاسبه امتیاز هر ستون به جاهایی که دو فضای خالی زیر هم قرار می‌گیرند، امتیازی تعلق نمی‌گیرد.

برای حل مسئله هم‌ترازی چندگانه، می‌توان همانند مسئله هم‌ترازی دوگانه، با استفاده از روش برنامه‌ریزی پویا(Dynamic programming) راه حل ارائه داد. در مسئله هم‌ترازی دوگانه، راه حل برنامه ریزی پویا را می‌توان به صورت زیر بازنویسی کرد:

$$V(i_1, i_2) \max_{(b_1,b_2)\in\{0,1\}^2-\{(0,0)\}} \{V(i_1 - b_1, i_2 - b_2) + \delta(S_1[i_1 b_1], S_2[i_2 b_2)\}$$

می‌توان به همین ترتیب برای مسئله هم‌ترازی چندگانه، فرض می‌کنیم k رشته داریم و آن‌ها را به صورت $S = S_1[1..n], S_2[1..n], ..., S_k[1..n_k]$ درنظر می‌گیریم. همچنین فرض می‌کنیم به ازای تمامی $j < k$ $S_j[0]$. بنابراین امتیاز هم‌ترازی چندگانه یا $V(i_1, i_2, ..., i_k)$ از معادله زیر به دست می‌آید:

$$V(i_1, i_2, ..., i_k) = \max_{(b_1,...,b_k)\in\{0,1\}^k-\{0^k\}} \{V(i_1 - b_1, ..., i_k - b_k) + SPscore(i_1 b_1, i_2 b_2, ..., i_k b_k)\}$$

که در آن $V(0, 0, ..., 0) = 0$ و

$$SPscore(i_1, i_2, ..., i_k) = \sum_{1<p<q<k} \delta(S[i_p], S[i_q])$$

بنابراین برای محاسبه امتیاز نهایی $V(n_1, n_2, ..., n_k)$ به روش برنامه‌ریزی پویا، جدول V را با استفاده از معادله بالا تکمیل کرده و با عملیات عقب گرد به رشته هم‌تراز شده می‌رسیم.

با توجه به این که $n_1 n_2 ... n_k$ ورودی داریم و هر ورودی 2^k حالت را بررسی می‌کند و k^2 حالت هم برای به‌دست آوردن امتیاز ستون، به صورت دو به دو محاسبه می‌شود، پیچیدگی زمانی این راه حل، $O(k^2 2^k n_1 n_2 ... n_k)$ است و همچنین پیچیدگی حافظه آن هم $O(n_1 n_2 ... n_k)$ است که پیچیدگی زمانی و حافظه بسیار بالایی دارد و از این رو مسئله هم‌ترازی چندگانه، یک مسئله np-hard است. بنابراین می‌توان این مسئله را در پیچیدگی زمانی کمتر و با الگوریتم‌های تقریبی حل نمود.

۴ روش ستاره مرکز (Center Star Method)

این روش یک الگوریتم تقریبی برای مسئله هم‌ترازی چندگانه است که امتیاز نهایی که محاسبه می‌گردد، حداکثر دوبرابر امتیاز هم‌ترازی بهینه است.

در این روش ابتدا فاصله بهینه $D(X, Y)$ را بین هر دو رشته X و Y محاسبه کرده و رشته‌ی S_c که عبارت $\sum_{i=1}^{k} D(S_c, S_i)$ را کمینه می‌کند، یعنی مجموع فواصل S_c تا سایر رشته‌ها کمینه باشد را به عنوان رشته مرکز درنظر می‌گیریم. سپس تمامی رشته‌های باقی مانده S_i را با رشته S_c هم‌تراز می‌کنیم. در نهایت برای هم‌ترازی همه این جفت رشته‌های هم‌تراز شده، تعدادی فضای خالی به هر رشته‌ای که لازم است اضافه می‌کنیم تا هم‌ترازی بین تمامی رشته‌ها حفظ گردد. شبه کد این روش در شکل ۲ مشاهده می‌شود.

مثالی از این روش را در شکل ۳ می‌توان مشاهده نمود. در این مثال پنج رشته DNA داریم و با فرض این که امتیاز missmatch/indel، یک در نظر گرفته شده، جدول فاصله، که در قسمت (b) مشاهده می‌شود را به دست می‌آوریم. سپس مجموع فواصل هر رشته تا سایر رشته‌ها را به دست

Center_Star_Method

Require: A set $\mathcal{S}$ of sequences

Ensure: A multiple alignment of M with sum of pair distances at most twice that of the optimal alignment of $\mathcal{S}$

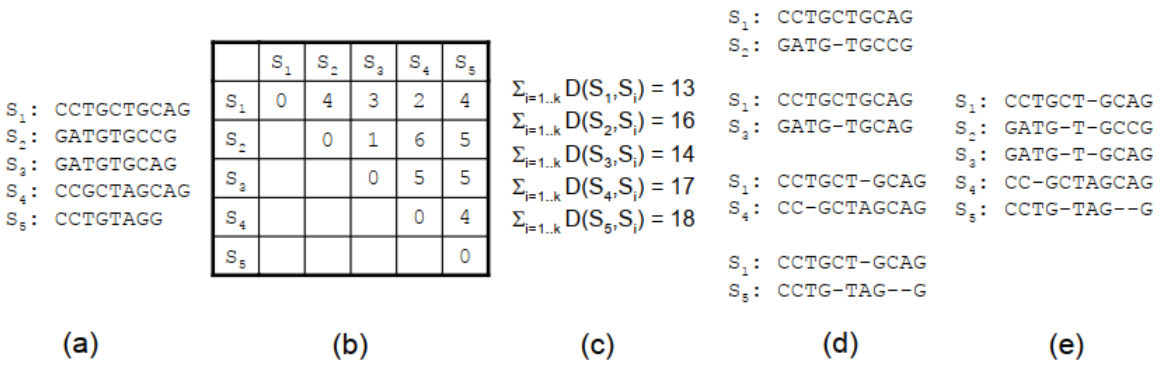
1: Find $D(S_i, S_j)$ for all i, j.

2: Find the center sequence S_c which minimizes $\sum_{i=1}^{k} D(S_c, S_i)$.

3: For every $S_i \in \mathcal{S} - \{S_c\}$, choose an optimal alignment between S_c and S_i.

4: Introduce spaces into S_c so that the multiple alignment $\mathcal{M}$ satisfies the alignments found in Step 3.

شکل ۲: روش ستاره مرکز



شکل ۳: مثال روش ستاره مرکز

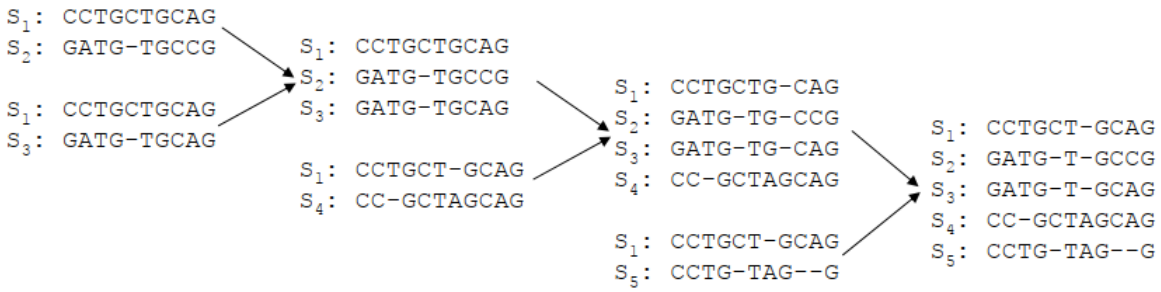
آورده و رشته‌ی S_1، که کمینه فاصله تا سایر رشته ها دارد را به عنوان رشته مرکز درنظر می‌گیریم. سپس در قسمت (d) بین رشته S_1 و سایر رشته‌ها هم‌ترازی دو گانه انجام می‌شود و درنهایت بین این هم‌ترازی‌های دوگانه اجتماع گرفته شده تا هم‌ترازی نهایی حاصل گردد.

برای پیدا کردن هم‌ترازی نهایی بین جفت رشته‌ها، در مرحله اول، دو جفت رشته را در نظر گرفته و با اضافه کردن فضای خالی در مکان‌های مناسب، این سه رشته را با هم هم‌تراز کرده و در مرحله بعد هم‌ترازی بین این سه رشته هم‌تراز شده و یک جفت رشته هم‌تراز شده دیگر را به دست می‌آوریم و این کار را ادامه می‌دهیم تا تمامی رشته‌ها هم‌تراز گردند. روش پیدا کردن اجتماع بین این رشته‌ها در مثال شکل ۴ آورده شده است.

می‌دانیم به ازای هر هم‌ترازی چندگانه $\mathcal{M}$، فاصله $d_{\mathcal{M}}()$ شرط نامساوی مثلث را برآورده می‌کند؛ یعنی به ازای هر سه رشته X، Y و Z، در یک هم‌ترازی چندگانه داریم:

$$d_{\mathcal{M}}(X, Y) \leq d_{\mathcal{M}}(X, Z) + d_{\mathcal{M}}(Z, Y)$$

بنابراین برای این که نشان دهیم فاصله هم ترازی به دست آمده از روش ستاره مرکز $(\mathcal{M})$ حداکثر دو برابر فاصله هم‌ترازی بهینه $(\mathcal{M}^*)$ است، با استفاده نامساوی مثلثی که در فوق توضیح داده شد، مجموع فواصل دو به دو رشته‌ها در $\mathcal{M}$ از روابط زیر به دست می‌آید:



شکل ۴: هم‌ترازی نهایی بین تمامی رشته‌ها در روش ستاره مرکز

$$
\begin{aligned}
&= \sum_{1 \le i \le j \le k} d_{\mathcal{M}}(i,j) \\
&= \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} d_{\mathcal{M}}(i,j) \\
&\le \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} [D(S_c,S_i) + D(S_c,S_j)] \\
&= \frac{k}{2} \sum_{i=1}^{k} D(S_c,S_i) + \frac{k}{2} \sum_{j=1}^{k} D(S_c,S_j) \\
&= k \sum_{j} D(S_c,S_j)
\end{aligned}
$$

و به همین ترتیب، با توجه به این که می‌دانیم مجموع فواصل S_c تا سایر رشته از مجموع فواصل هر رشته S_i تا سایر رشته کوچک‌تر یا مساوی است، می‌توان نشان داد:

$$
\begin{aligned}
&= \sum_{1 \le i \le j \le k} d_{\mathcal{M}^*}(i,j) \\
&\ge \sum_{1 \le i \le j \le k} D(S_i,S_j) \\
&= \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} D(S_i,S_j) \\
&\ge \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} D(S_c,S_j) \\
&= \frac{k}{2} \sum_{j=1}^{k} D(S_c,S_j)
\end{aligned}
$$

که در آن مجموع فواصل در $\mathcal{M}^*$ حداقل $\frac{k}{2} \sum_j D(S_c,S_j)$ است. بنابراین باتوجه به عبارات به دست آمده می‌بینیم که مجموع فواصل $\mathcal{M}$ حداکثر دو برابر مجموع فواصل بهینه برای مجموعه رشته S است.

لازم به ذکر است که پیچیدگی زمانی این روش، $O(k^2n^2)$ است، زیرا در مرحله اول برای محاسبه فاصله بهینه بین دو به دوی رشته‌ها $O(k^2n^2)$ زمان لازم است. در مرحله دوم برای پیدا کردن رشته مرکز، پیچیدگی زمانی $O(k^2)$ داریم. در مرحله سوم برای هم‌ترازی بین رشته مرکز و سایر رشته‌ها به $O(kn^2)$ زمان نیاز است و در نهایت برای اجتماع‌گیری بین تمام جفت رشته‌های هم‌تراز شده نیز پیچیدگی زمانی $O(kn^2)$ داریم.

۵ روش‌های پیش‌برنده

روش‌های پیش‌برنده به طور کلی شامل سه مرحله می‌شوند:

۱. به ازای هر جفت دنباله، فاصله دو به دو بین آن‌ها محاسبه می‌شود.

۲. از روی ماتریس به دست آمده یک درخت فیلوژنی ایجاد می‌شود به‌طوری که بدانیم دنباله‌های مشابه، در درخت نیز نزدیک به هم قرار دارند.

۳. با استفاده از این درخت هم‌ترازی بین رشته‌ها انجام می‌شود.

یکی از روش‌های پیش‌برنده، روش ClustalW است که در ادامه به آن می‌پردازیم.

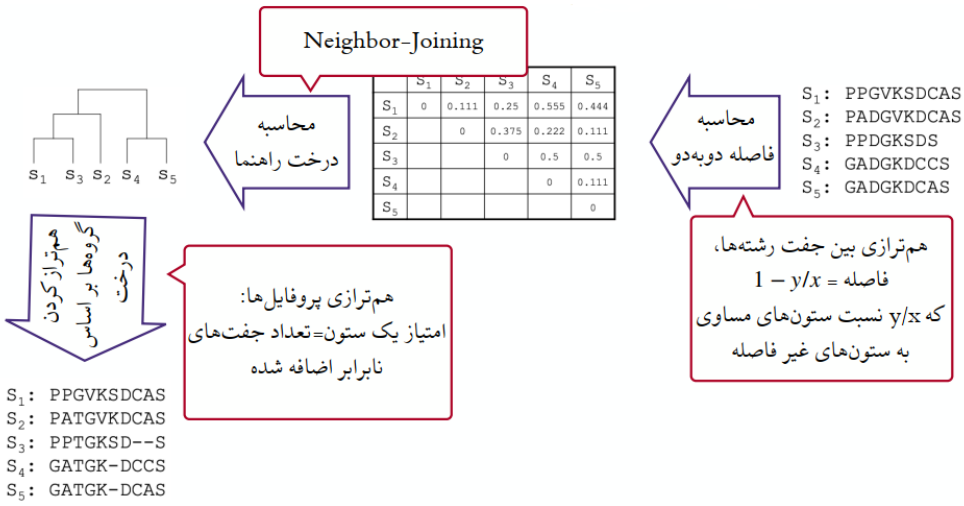
۱.۵ ClustalW

در این روش، ابتدا هم‌ترازی بین تمامی جفت رشته‌ها انجام می‌شود و بر اساس آن، فاصله بین هر دو جفت رشته از رابطه $1 - \frac{y}{x}$ به دست می‌آید که x تعداد ستون‌های بدون فاصله و y ستون‌های مساوی هستند. به عنوان مثال اگر دو رشته زیر را در نظر بگیریم:

S1: PPGVKSDCAS
S2: PPDGKSD - - S

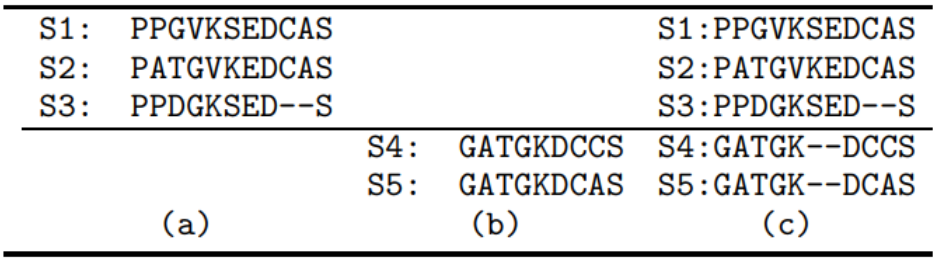
۸ ستون بدون فاصله و ۶ ستون مساوی داریم؛ از این رو فاصله برابر است با: $1 - \frac{6}{8} = 0.25$.

در مرحله بعدی، این الگوریتم یک درخت با استفاده از روش neighbour joining ایجاد می‌کند و درنهایت با توجه به درخت ایجاد شده، گروه‌ها را با استفاده از روش هم‌ترازی پروفایل‌ها، هم‌تراز می‌کند؛ به این صورت که در هر مرحله دو گروه هم‌تراز شده را درنظر گرفته و به اعضای هر گروه در مکان‌های مشخص و یکتا برای اعضای آن گروه تعدادی فاصله اضافه می‌شود تا این دو گروه با هم هم‌تراز شده و در عین حال هم‌ترازی بین رشته‌های یک گروه نیز بدون تغییر باقی بماند. در شکل ۵ مثالی برای الگوریتم ClustalW را مشاهده می‌کنید.



شکل ۵: مثال روش ClustalW

همچنین مثالی برای هم‌ترازی پروفایل بین دو مجموعه رشته در شکل ۶ نشان داده شده است. در این مثال، برای ایجاد هم‌ترازی بهینه بین دو مجموعه رشته که در قسمت (a) و (b) دیده می‌شود، به ستون ۵ و ۶ مجموعه دوم، فاصله اضافه شده است تا این هم‌ترازی به درستی انجام شود.



شکل ۶: مثال هم‌ترازی پروفایل در ClustalW

۶ روش‌های تکراری

در این روش‌ها، ابتدا یک هم‌ترازی چندگانه اولیه ایجاد شده سپس به صورت تکرارشونده، تلاش بر بهبود این هم‌ترازی است. یکی از روش‌های تکراری، روش MUSCLE است که در ادامه به آن می‌پردازیم.

۱.۶ روش MUSCLE

این روش در مرحله اول، از یک روش پیش‌رونده مثل ClustalW برای محاسبه فواصل بین رشته‌ها و ایجاد یک درخت راهنما استفاده می‌کند با این تفاوت که برای بهبود دقت، هنگامی که دو پروفایل هم‌تراز شدند، از امتیاز لگاریتم امید استفاده می‌کند تا تعداد فواصل را کاهش دهد. این امتیاز به صورت زیر محاسبه می‌شود:

$$LE(A_1[i], A_2[j]) = (1 - f_G^i)(1 - f_G^j) log \sum_{x,y \in A} f_x^i f_y^j \delta(x, y)$$

همچنین برای بهبود کارایی آن، برای محاسبه فاصله، از فاصله k-mer استفاده می‌کند به این صورت که همه زیر رشته‌های k تایی که در هر دو دنباله مشترک هستند را در نظر می‌گیرد و برای ایجاد درخت راهنما، از روش UPGMA استفاده می‌کند که سرعت بالاتری نسبت به روش neighbour joining برای ایجاد درخت دارد.

در مرحله دوم با استفاده از هم‌ترازی چندگانه ایجاد شده در مرحله اول، یک پیش‌روی ارتقا دهنده اعمال می‌شود؛ به این صورت که به ازای دنباله‌های هم‌تراز شده، یک فاصله دو به دوی جدید تعریف می‌شود که به آن فاصله Kimura هم می‌گویند و این فاصله به ازای هر دو گونه به صورت $-ln(1 - D - D^2/5)$ محاسبه می‌شود. در این عبارت D بازهای یکتایی است که در هر دو رشته با هم هم‌تراز شده‌اند. در نهایت، بر اساس این ماتریس فاصله، درخت راهنما مجدد ساخته شده و دنباله‌ها مجددا هم‌تراز می‌شوند.

در مرحله سوم بر روی این هم‌ترازی چندگانه اصلاحاتی انجام می‌شود به این صورت که با حذف هر یال، درخت به دو زیر درخت تقسیم شده و دو زیر مجموعه از دنباله‌ها، متناظر با هر زیر درخت خواهیم داشت. هم‌ترازی چندگانه هر کدام از این دو زیر مجموعه را می‌توان با استفاده از هم‌ترازی که در دور قبل محاسبه شده است، به دست آورد. حال با استفاده از روش هم‌ترازی پروفایل، این دو مجموعه رشته را مجددا هم‌تراز می‌کنیم. اگر امتیاز این هم‌ترازی چندگانه بهبود یافت، این هم‌ترازی نگه داشته می‌شود.