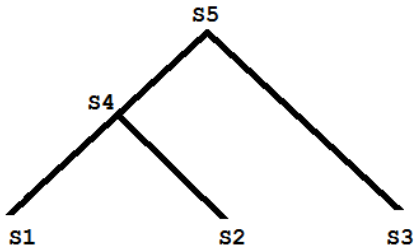
در این جلسه قصد داریم رشته‌ها را با استفاده از روش فاصله برحسب درخت تکامل هم‌ردیف کنیم. تا این‌جا درخت را با استفاده از یک‌سری اطلاعات ساخته‌ایم، هدف آن است که روی ساختار درخت alignment انجام دهیم. به طور مثال فرض کنید سه رشته ورودی $S_1$، $S_2$ و $S_3$ و ساختار درخت را طبق شکل۱ در اختیار داریم و به دنبال پیدا کردن بهترین هم‌ردیفی این سه رشته هستیم. در این حالت باید برای گره‌های داخلی درخت را تعدادی رشته درنظر گرفته ( نظیر $S_4$ و $S_5$) و هم‌ردیفی سازگار با درخت را بدست آوریم. به عنوان مثال برای به‌دست آوردن هم‌ردیفی سازگار با درخت موجود در شکل۱ نودهای $S_4$ و $S_5$ باید طوری انتخاب شوند که امتیازهای $d(S_1,S_4)$ و $d(S_2,S_4)$ و $d(S_5,S_4)$ و $d(S_5,S_3)$ امتیازهای بهینه باشند.




شکل۱. ساختار درخت با ۳ رشته ورودی.

درنهایت با داشتن نودهای داخلی می‌توان امتیاز هم‌ردیفی را به‌صورت زیر محاسبه کرد:

$$\sum_{(u,v)E(T)} d(S_u, S_v) \quad (1)$$

بنابراین هدف آن است تا رشته‌ها در گره‌های داخلی را طوری انتخاب کنیم که امتیاز فوق مینمم شود. اثبات شده است این مسئله یک مسئله NP-complete است، بنابراین برای حل آن از روش‌های تقریبی استفاده می‌کنیم. در این‌جا به بررسی تعدادی از روش‌های تقریبی خواهیم پرداخت.

## ۱ روش Lifted tree

### ۱.۱ تکنیک اول:

در این روش با در اختیار داشتن ساختار درخت $T$ و رشته‌های $S = \{S_1, S_2, \dots, S_k\}$، از پایین به بالا حرکت کرده و هر رشته در گره داخلی را با یکی از رشته‌های داخل $S$ طوری جایگزین می‌کنیم که طبق معادله زیر بهترین امتیاز را داشته باشد.

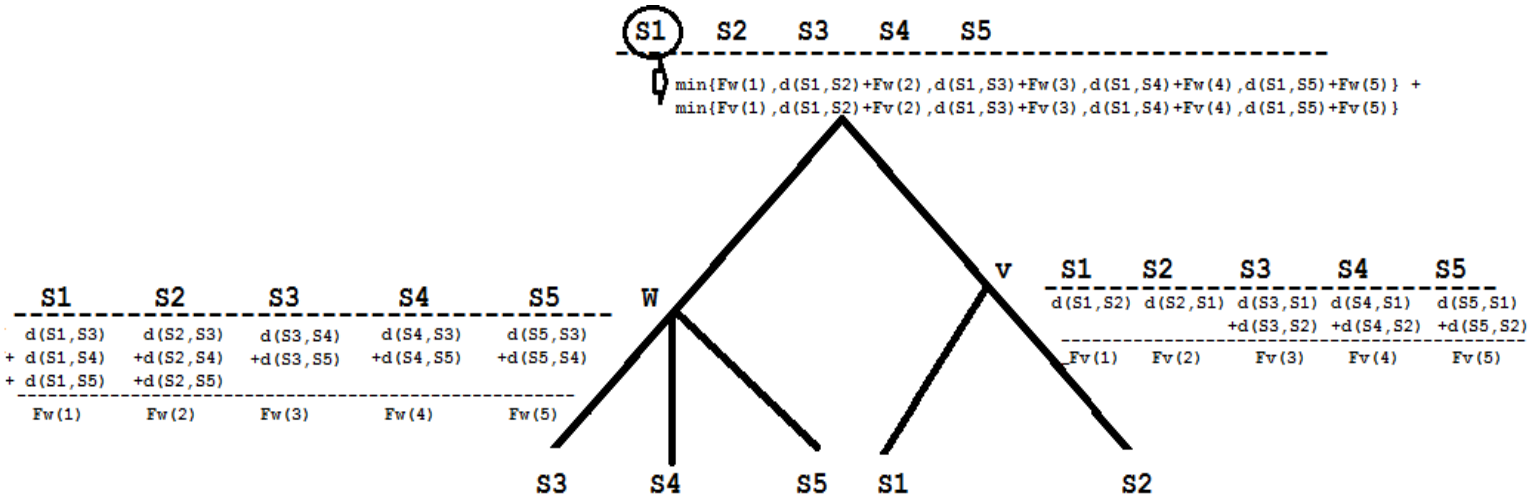
$$F_v(x) = \sum_{v'} \min_{x'} \big(d(x', x) + F_{v'}(x')\big) \quad (2)$$



نگارنده: عادله بیطرفان

ایمیل: adele.bitarafan@gmail.com

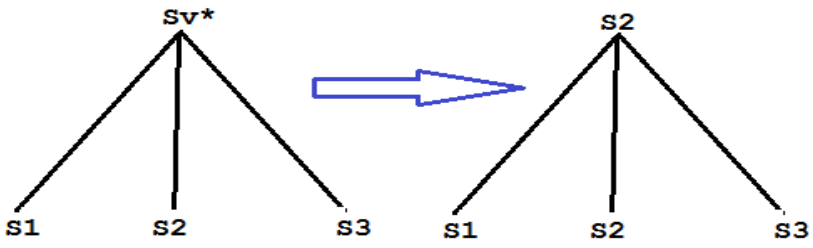
در معادله فوق $v'$ تمام فرزندان نود $v$ است و $F_v(x)$ فاصله برای زمانی است که $x$ را در گره $v$ قرار دهیم. به عنوان مثال، طبق شکل۲ فرض کنید ۵ رشته ورودی و ساختار درخت را داریم و هدف آن است رشته‌های داخلی را از رشته‌های ورودی انتخاب کنیم. در شکل۲ امتیاز به ازای قرار دادن هر رشته در هر گره محاسبه شده است که در هر گره رشته‌ای را انتخاب می‌کنیم که بهترین امتیاز را داشته باشد.




شکل۲. محاسبه امتیازهای بهینه برای هر گره.

### ۱.۲ تکنیک دوم:

در این روش از روی درخت بهینه $T^*$ درخت $T^L$(Lifted tree) را می‌سازیم. بدین صورت که برای ساخت درخت$T^L$ از پایین به بالا حرکت کرده و هر رشته در گره داخلی را با رشته‌ای از برگ‌ها جایگزین می‌کنیم که کم‌ترین فاصله را با بهترین گره داخلی از درخت $T^*$ دارد. به عنوان مثال در شکل زیر رشته $S_2$ به عنوان رشته در گره داخلی درخت $T^L$ انتخاب شده است چرا که داریم :



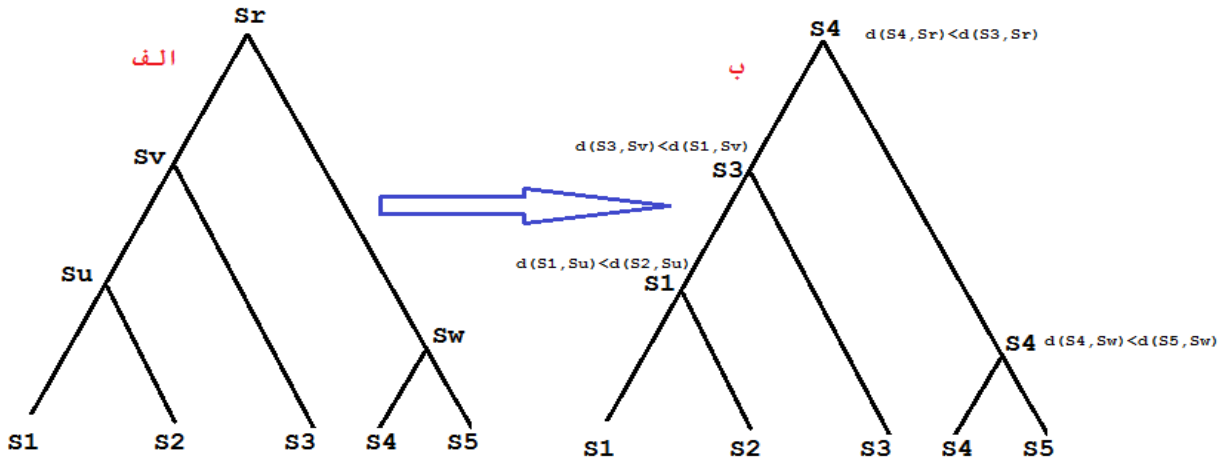

$$d(S_2, S_v^*) \leq min(d(S_1, S_v^*), d(S_3, S_v^*))$$



نگارنده: عادله بیطرفان

ایمیل: adele.bitarafan@gmail.com

با توجه به توضیحات فوق، درخت lifted tree حاصل از درخت optimal در شکل ۳(الف) به صورت شکل ۳(ب) خواهد بود.

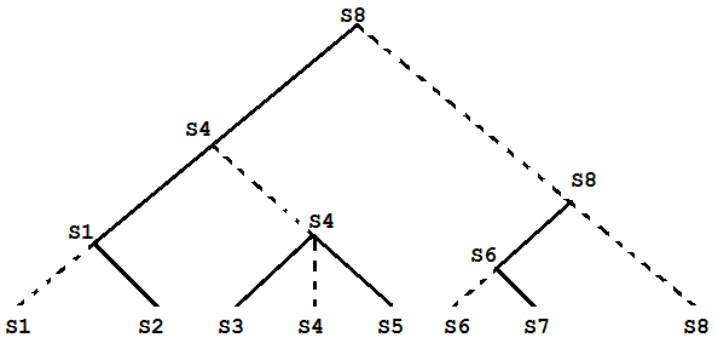



شکل۳. ساخت درخت Lifted tree (ب) از روی درخت Optimal (الف).

مشخص است که درخت ساخته شده با استفاده از تکنیک اول امتیاز بیش‌تری ( وضع بدتری) نسبت به درخت ساخته شده با استفاده از تکنیک دوم دارد و به همین‌ترتیب نیز درخت ساخته شده با استفاده از تکنیک دوم نسبت به درخت optimal امتیاز بیش‌تری دارد.

**قضیه:** همردیفی $T^L$ دارای امتیاز مساوی و یا کم‌تر از دو برابر همردیفی بهینه $T^*$ می‌باشد.

**اثبات:** فرض کنید شکل۴ یک درخت lifted tree باشد. خط‌های نقطه‌چین مسیر lifted را نشان می‌دهند و هر یال ساده یک مسیر نقطه‌چین به برگ‌ها دارد.




شکل۴. درخت Lifted tree. خط‌های نقطه‌چین مسیر Lifted را نشان می‌دهد.



نگارنده: عادله بیطرفان

ایمیل: adele.bitarafan@gmail.com

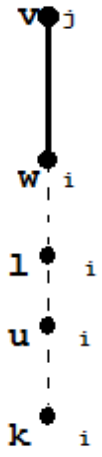
برای درخت lifted باید امتیاز همه یال‌ها را حساب کنیم. طبق شکل۴ یال‌های ساده معادل یک مسیر هستند که در این مسیر تنها یال ساده distance دارد و یال‌های دیگر دارای distance صفر می‌باشند، بنابراین برای هر یال ساده $(v,w)$ یک مسیر نقطه‌چین تا برگ‌ها داریم که تحت lifted tree امتیاز این مسیر برابر است با $d(S_v,S_w)$ . بنابراین امتیاز کل برای درخت lifted tree برابر است با جمع تمام مسیر یال‌های ساده تا برگ، درنتیجه تنها کافی است امتیاز یال‌های ساده را به‌دست آوریم.

بنابراین اگر ابتدا به یک یال $(v,w)$ نگاه کنیم که $S_i$ به $w$ و $S_j$ به $v$ تخصیص یافته باشد، آنگاه داریم:


$$d(S_j,S_i) \le d(S_j,S_v^*) + d(S_v^*,S_i) \le ۲d(S_i,S_v^*) \qquad (۳)$$

در رابطه فوق نامساوی اول به‌دلیل نامساوی مثلثی و نامساوی دوم به‌دلیل نحوه انتخاب در درخت $T^L$ است، چرا که فرض شده است گره $v$، lifted شده است.

برای درخت optimal نیز همین مسیر را حساب می‌کنیم. اگر مسیر گره $w$ را تا برگ $k$ ادامه دهیم همه گره‌های داخلی به $S_i$ به بالا آمده اند که با توجه به نامساوی مثلثی در درخت optimal داریم:




$$d(S_i,S_v^*) \le \sum_{\substack{(u,v)\in E(T)\\(u,v)\in P}} d(S_u^*,S_v^*) \qquad (۴)$$

با توجه به دو رابطه (۳) و (۴) داریم:

$$d(S_i,S_v^*) \ge \frac{1}{2}\, d(S_i,S_j) \qquad (5)$$



نگارنده: عادله بیطرفان

ایمیل: adele.bitarafan@gmail.com

با در نظر گرفتن مسیر P رابطه فوق می‌گوید:

مجموع امتیازات در مسیر P × 2 ≤ مجموع امتیازات Lifted در مسیر P

بنابراین قضیه اثبات می‌شود.

تا این‌جا توانستیم multi alignment را با استفاده از درخت به‌دست آوریم. حال فرض کنید درخت را در اختیار نداریم، در این حالت از روش‌های هیورستیک استفاده می‌کنیم. یکی از این روش‌ها، روش‌های هم‌ردیفی Progressive اند که در این‌جا تنها به الگوریتم clustal W اشاره می‌کنیم.

## ۲ الگوریتم Clustal W

در این روش ابتدا دو رشته انتخاب و با یکدیگر هم‌ردیف می‌گردند. سپس رشته سوم به آن‌ها اضافه شده و همین‌طور مسئله پیش می‌رود. مراحل کلی این الگوریتم به‌صورت زیر است:

۱. با هم‌ردیفی دوتایی تمامی زوج‌ها، distance دودویی همه رشته‌ها را بدست می‌آوریم ( به عبارت دیگر ماتریس فاصله را محاسبه می‌کنیم).
۲. با استفاده از الگوریتم neighbor joining یک درخت می‌سازیم.
۳. حال در درخت ساخته شده از پایین به بالا حرکت کرده و هنگام برخورد به یک گره داخلی رشته آن را با استفاده از روش هم‌ردیفی پروفایل به پروفایل انتخاب می‌کنیم.

**مثال:**

فرض کنید ۵ رشته $\{S_1, S_2, S_3, S_4, S_5\}$ را به‌صورت زیر در اختیار داریم.

$S_1 = P\ P\ G\ V\ K\ S\ D\ C\ A\ S$ $S_2 = P\ A\ D\ G\ V\ K\ D\ C\ A\ S$ $S_3 = P\ P\ D\ G\ K\ S\ D\ S$

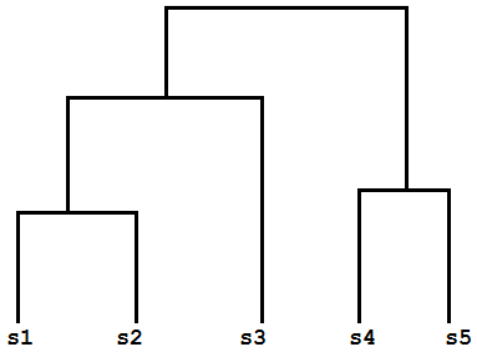
$S_4 = G\ A\ D\ G\ K\ D\ C\ C\ S$ $S_5 = G\ A\ D\ G\ K\ D\ C\ A\ S$

برای به‌دست آوردن multi alignment آن‌ها با استفاده از روش clustalW ابتدا ماتریس فاصله آن‌ها را محاسبه کرده، سپس با استفاده از ماتریس فاصله و الگوریتم neighbor joining درخت آن را می‌سازیم. فرض کنید درخت حاصل همانند شکل زیر باشد.



نگارنده: عادله بیطرفان

ایمیل: adele.bitarafan@gmail.com




با توجه به درخت فوق، ابتدا $S_1$ را با $S_2$ هم‌ردیف کرده سپس مجموع را با $S_3$ هم‌ردیف می‌کنیم. از طرف دیگر $S_4$ را با $S_5$ هم‌ردیف نموده و سپس دو هم‌ردیفی خاصل را با یک‌دیگر هم‌ردیف می‌کنیم. در این حالت نتایج حاصل از هم‌ردیفی به صورت زیر خواهد شد:

$S_1 = P - P\ G\ V\ K\ S\ D\ C\ A\ S$

$S_2 = P\ A\ D\ G\ V\ K - D\ C\ A\ S$

$S_3 = P\ P\ D\ G - K\ S\ D - - S$

$S_4 = G\ A\ D\ G - K - D\ C\ C\ S$

$S_5 = G\ A\ D\ G - K - D\ C\ A\ S$

که برای امتیازدهی بین دو پروفایل در clustalW از معادله زیر استفاده می‌کنیم.

$$d(A_1[i],\ A_2[j]) = \sum_{x,y} g_x^i \times g_y^j \times \sigma(x,y) \qquad (6)$$

در معادله فوق $g_x^i$ تعداد آمینواسیدهای $x$ است که در ستون $i$ مشاهده شده است.



نگارنده: عادله بیطرفان

ایمیل: adele.bitarafan@gmail.com