۱ Multiple sequence alignment

مسئله‌ی همردیفی چندین رشته را می‌توان به صورت ریاضی زیر تعریف نمود:

K رشته‌ی $S_1, S_2, \dots, S_K$ را در اختیار داریم و می‌خواهیم آن‌ها را با یکدیگر همردیف نماییم. این همردیفی بدان معناست که با قرار دادن تعدادی gap میان رشته‌ها آن‌ها را هم‌اندازه کرد و با تطبیق ستون‌ها، همردیفی را ایجاد نماییم.

مثال: رشته‌های زیر را در نظر می‌گیریم:

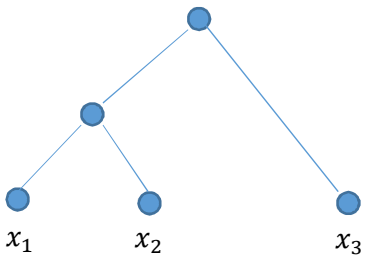
$$X_1 = ACCG \quad X_2 = ACA \quad X_3 = ACGA$$

دارای همردیفی‌های زیر می‌باشد:

ACCG	ACCG_
AC_A	A_C_A
ACGA	A_CGA

هدف آن است که از بین تمام همردیفی‌های ممکن، آنی را بیابیم که دارای بیش‌ترین امتیاز باشد. برای امتیازدعی نیاز به جدول امتیازدهی می‌باشد که به طور معمول از مدل‌های تکاملی به دست می‌آید.

به طور مثال اگر رابطه‌ی تکاملی رشته‌ها را بدانیم و مدل تکاملی را در اختیار داشته باشیم، می‌توانیم مانند همردیفی دو رشته‌ای، ماتریس امتیاز را محاسبه نماییم. ولی به طور معمول چنین چیزی را در اختیار و همین مسئله را مشکل می‌نماید.



نکته: حتی با داشتن ماتریس امتیازدهی، محاسبه‌ی همردیفی بهینه کار دشواری است. البته می‌توان از DP استفاده نمود و سپس برای بدست آوردن امتیاز یک خانخ از جدول، نیاز به بیشینه نمودن از $2^k - 1$ عدد می‌باشد و در مجموع $\prod_{i=1}^{k} n_i$ نیز باید پر گردد که در آن‌ها n_i طول رشته‌ی iام می‌باشد. در نتیجه میزان محاسبات $O(2^k n^k)$ خواهد بود.

بنابراین تنها راه باقی‌مانده آن است که به طور $Heuristic$ امتیازدهی نماییم و بهترین همردیفی را محاسبه کنیم.

چندین روش امتیازدهی به کار گرفته شده است که به سه نمونه از آن‌ها می‌پردازیم:

نگارنده: محمدجواد رضایی سراجی
ایمیل: mohammad.jrs@gmail.com

۱. مجموع جفت‌ها
۲. فاصله تا مرجع
۳. فاصله بر پایه‌ی درخت تکامل

۱-۱ مجموع جفت‌ها

ماتریس امتیازدهی در این روش امتیازدهی به صورت زیر می‌باشد:

$$S_{sp} = \sum_{i<j} d(x_i^*, x_j^*)$$

یعنی پس از همردیفی هر دو رشته را در نظر گرفته و با توجه به ماتریس امتیازدهی دوتایی، امتیاز مربوطه را محاسبه کرده و سپس همه‌ی آن‌ها را با یکدیگر جمع می‌نماییم. بدلیل این که محاسبه‌ی امتیازات در هر ستون مستقل از بقیه ستون‌ها است می‌توانیم از DP استفاده نموده و مسئله را حل کنیم اما همانطور که بیان شد پیچیدگی محاسبات بسیار بالایی دارد.

الگوریتم MSA سعی می‌کند که با کنار گذاشتن تعدادی مسیرها، سریع‌تر به جواب بهینه برسد.

۱-۲ فاصله تا مرجع

در این روش یکی از رشته‌ها را به عنوان مرجع انتخاب نموده و امتیاز را از رابطه‌ی زیر محاسبه می‌نماییم:

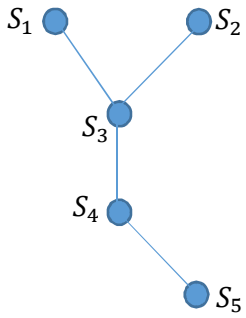
$$S_c = \sum_{i=1}^{k} d(x_i^*, c)$$

که در آن c رشته‌ی مرجع می‌باشد.

بدیهی است که در صورت داشتن c بهتر است که هر یک از رشته‌ها را به طور مجزا با c همردیف کنیم و در صورت نیاز gap ایجاد نماییم. اما پیدا کردن c کار ساده‌ای نمی‌باشد.

به دلیل اینکه به این موضوع در آینده نیاز خواهیم داشت، تعریف زیر را بیان می‌کنیم:

تعریف: قرار می‌دهیم $S = \{S_1, \dots, S_k\}$ و T یک درختی که هر گره‌ی آن با یک رشته در S برچسب خورده است. به یک همردیفی S می‌گوییم سازگار با T است اگر امتیاز هر دو رشته‌ی مجاور در درخت در همردیفی S، امتیاز بهینه بین دو رشته را داشته باشند.



$$D(S_3, S_4) = d(X_3, X_4)$$

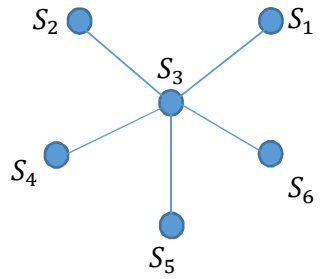
pairwise induced

قضیه: برای هر S و T می‌توان به طور بهینه هردیفی سازگار با T را یافت.

اثبات. به سادگی می‌توان پیش رفت و یکی یکی رشته‌ها را همردیف نمود. زمان مورد نیاز نیز $O(kn^2)$ می‌باشد.

۱-۲-۱ روش $Center-star$

در این روش، یکی از رشته‌ها را به عنوان مرجع انتخاب نموده و همردیفی سازکار با درخت ستاره را بدست می‌آوریم. از میان تمام رشته‌ها، آن مرجعی که بیش‌ترین امتیاز را دارد را به عنوان جواب انتخاب می‌کنیم. با محاسبه‌ی تمام همردیفی‌های دوتایی و با انجام $O(k^2n^2)$ عملیات می‌توان جواب مسئله را یافت.



در حقیقت روش $Cemter-star$ یک روش تقریبی برای محاسبه‌ی مقدار بهینه‌ی امتیاز مجموع جفت‌ها می‌باشد. بدین منظور می‌بایست فرض نامساوی مثلثی را در ماتریس امتیازدهی داشته باشیم. یعنی:

$$\sigma(x,y) \le \sigma(x,z) + \sigma(z,y)$$

در پاره‌ای از مواقع این نامساوی برقرار نمی‌باشد و در نتیجه نتایج نیز نادرست خواهد بود.

لم. با فرض برقرار بودن نامساوی مثلثی، برای هر همردیفی $Center-star$ میان رشته‌ها داریم:

$$d(S_i,S_j) \le d(S_i,S_c) + d(S_c,S_j) \le D(S_i,S_c) + D(S_c,S_j)$$

قضیه. $\frac{S_c^*}{S_{sp}^*} \le 2\left(1-\frac{1}{k}\right) < 2$ که در آن S_{sp}^* امتیاز همردیفی بهینه در روش امتیازدهی مجموع جفت‌ها بوده و S_c^* امتیاز همردیفی بهینه فاصله تا مرجع برای هر $\exists c \in S$.

اثبات. برای همردیفی بهیمه خواهیم داشت:

$$S_{sp}^* = \frac{1}{2}\sum_{(i,j)} d^*(S_i,S_j)$$

$$\ge \frac{1}{2}\sum_{(i,j)} D(S_i,S_j) \ge \frac{1}{2}k\sum_{j} D(S_c,S_j)$$

آخرین نامساوی از این نکته نتیجه می‌شود که رشته‌ی مرکزی را به گونه‌ای انتخاب کردیم که رابطه‌ی زیر همواره برقرار است:

$$\sum_j D(S_c,S_j) \le D(S_c',S_j)\ \forall S_c' \in S$$

از طرف دیگر داریم:

$$S_c^* = \frac{1}{2}\sum_{(i,j)} d(S_i, S_j)$$

$$\leq \frac{1}{2}\sum_{(i,j)} \left[D(S_i, S_c) + D(S_c, S_j)\right]$$

$$= (k-1)\sum_{j} D(S_c, S_j)$$

نامساوی اول از لم قبلی و مساوی دوم به این خاط است که هر $D(S_c, S_j)$ به تعداد $2(k-1)$ بار در رابطه دیده شده است.
در نتیجه داریم:

$$\frac{S_c^*}{S_{sp}^*} \leq \frac{k-1}{\frac{1}{2}k} = 2\left(1 - \frac{1}{k}\right) < 2$$