# ۱ مقایسه‌ی دو رشته‌ی DNA

با دانستن آن‌که موجودات چگونه متحول شده و تکامل یافته‌اند و داشتن مدل تکاملی مناسب، برای هر دو رشته‌ی $x^n$ و $y^n$، می‌توان احتمال وقوع آن‌ها را تحت فرض H محاسبه نمود:

$$L = \log(P(x^n, y^n|H))$$

که به عنوان مثال در مدل $Jukes - Cantor$، بازها مستقل از یکدیگر بوده و در نتیجه:

$$L = \sum_{i=1}^{i=n} \log(P(x_i, y_i|H)$$

از طرف دیگر به علت آن‌که مدل ماکوف در نظر گرفته شده است و برگشت‌پذیر در زمان است، می‌توان $x_i$ را به عنوان پدر در نظر گرفت و در نتیجه داریم:

$$L = \sum_{i=1}^{i=n} \log(P(x_i \to y_i|H)$$

بنابراین کافی است که ماتریس جابجایی را در اختیار داشته باشیم ( به زمان $t$ وابسته است )، تا بتوانیم میزان $L$ را محاسبه نماییم.

Substituation Matrix

| | A | C | G | T |
|---|---|---|---|---|
| A | | | | |
| C | | | | |
| G | | | | |
| T | | | | |

$\log(P(G \to C|H)$

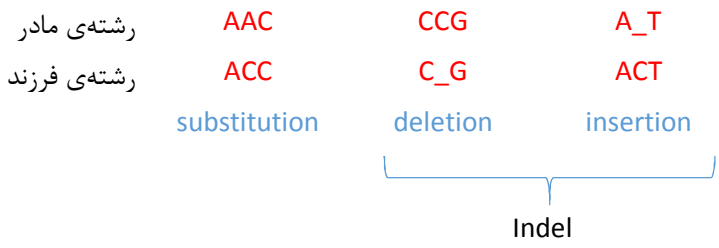
نکته: پارامترهای مدل را می‌توان در صورت عدم تخمین قبلی از روی $L$ تخمین زد. به عنوان مثال اگر $t$ را در اختیار نداشته باشیم، با محاسبه‌ی $L(t)$ و مشتق‌گیری از آن، می‌توان $t$ بهینه را بدست آورد.

## ۱-۱ مروری بر جهش‌ها

به عنوان مثال، هنگامی که دو رشته‌ی مربوط به $\beta - \text{globin}$ را در انسان و موش مقایسه می‌کنیم، مشاهده می‌شود که این دو رشته از نظر طول با یکدیگر یکی نیستند. در نتیجه مدل‌هایی نیاز داریم که اتفاقات جهشی دیگر را مانند $indel$ ها را مدل نماید.

جهش‌های مورد مطالعه:

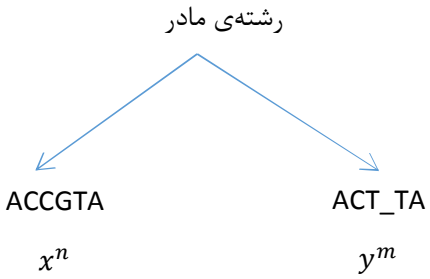
نگارنده: محمدجواد رضایی سراجی
ایمیل: mohammad.jrs@gmail.com




نکته: زمانی که رشته‌ی مادر ناشناخته است، تفاوتی میان حذف و اضافه وجود ندارد و در این حالت به آن‌ها $indel$ اطلاق می‌شود.

حال سوال این است که برای دو رشته‌ی $x^n$ و $y^m$ که دارای طول‌های متفاوت می‌باشند، چگونه می‌توان مقدار $L = \log(P(x^n, y^m|H))$ را محاسبه نمود. ابتدا باید دید که مدل $H$ چگونه است.

یک مدل ساده که می‌توان برای $Indel$ در نظر گرفته است، این است که این اتفاقات مستقل از بقیه اتفاقات بوده و در نتیجه خواهیم داشت:




و در نتیجه:

$$L = \log\big(P(A \to A)\big) + \log\big(P(C \to C)\big) + \log\big(P(C \to T)\big) + \log\big(P(G \to -)\big) + \log\big(P(T \to T)\big) + \log\big(P(A \to A)\big)$$

با وجود چنین مدل ساده‌ای، هنوز نیاز داریم که $x^n$ و $y^m$ را با هم همردیف کنیم و برای هر همردیف کردن یک $L$ بدست می‌آوریم. برای محاسبه‌ی احتمال کلی مجبود بوده تا روی تمام همردیفی‌ها جمع ببندیم. یعنی:

$$L = \sum_{\mathcal{A}} \log\big(P(x^n \to y^m|H, \vartheta).P(\vartheta|H)\big)$$

که در آن $\mathcal{A}$ یک همردیفی میان $x^n$ و $y^m$ می‌باشد.

مسئله‌ی دیگر آن است که همردیفی $\mathcal{A}^*$ را بیابیم به طوری که در آن:

$$P(x^n \to y^m|H, \mathcal{A}^*) \geq P(x^n \to y^m|H, \mathcal{A})$$

برای تمام همردیفی‌های دیگر باشد. یعنی به دنبال تخمین همردیفی از روی داده بوده و قصد محاسبه‌ی احتمال را نداریم.

مسئله‌ی دیگری که مطرح می‌باشد این است که تشخیص دهیم که آیا دو رشته‌ی $x^n$ و $y^m$ دارای جد مشترک هستند یا خیر. به عبارت دیگر در این حالت دو فرض پیش رو داریم:

- فرض $H_1$ که در آن فرض می‌شود که $x^n$ و $y^m$ تحت مدل تکاملی و با هم تولید گردیده‌اند.
- فرض $H_2$ که در آن فرض می‌شود که این دو رشته مستقل از یکدیگر می‌باشند.

می‌دانیم که در آزمون فرض، بهترین تصمیم‌گیری بر پایه نسبت احتمالات گرفته می‌شود، یعنی:

$$S = \log \frac{P(x^n, y^m|H_1)}{P(x^n, y^m|H_2)} \qquad (*)$$

به هرحال اگر همریفی را بدانیم، آن‌گاه دو رشته با هم دارای طول مشترک $K$ گردیده و داریم:

$$S_{\mathcal{A}} = \sum_{k=1}^{K} \frac{P(\tilde{x}_k, \tilde{y}_k|H_1)}{P(\tilde{x}_k|H_2).P(\tilde{y}_k|H_2)}$$

حال دو راه وجود دارد، یکی آن‌که $\mathcal{A}^*$ را بیابیم به طوری که $S_{\mathcal{A}^*} \geq S_{\mathcal{A}}$ برای تمام همردیفی‌های $\mathcal{A}$. در این حالت $\mathcal{A}^*$ همردیفی‌ای می‌باشد که در آن بیش‌ترین امتیاز رابطه وجود دارد و اگر بتواند بر یک آستانه‌ای غلبه کند، قبول می‌کنیم که دو رشته از یک جا آمده‌اند.

راه دوم آن است که $S$ در $(*)$ را بر روی تمام همردیفی‌ها جمع ببندیم، یعنی:

$$S = \log \frac{\sum P(x^n, y^m|H_1, \mathcal{A})}{\sum P(x^n, y^m|H_2, \mathcal{A})}$$

با مدل ساده‌ای که برای $indel$ ها در نظر گرفتیم و برای یافتن همردیفی بهینه، چه در حالت محاسبه‌ی احتمالات و چه مقایسه‌ی دو رشته به ماتریس زیر نیاز داریم:

| | A | C | G | T | _ |
|---|---|---|---|---|---|
| A | | | | | |
| C | | | | | |
| G | | | | | |
| T | | | | | |
| _ | | | | | |

$$l(c, -) = \log\big(P(C \to -)\big)$$

$$\sigma(C, -) = \log \frac{P(C \to -)}{P(C).P(-)}$$

$$l(G, C) = \log\big(P(G \to C)\big)$$

$$\sigma(G, C) = \log \frac{P(G \to C)}{P(G).P(C)}$$

که در آن به $\sigma(\mathrm{X}, -)$ و $\sigma(-, X)$ پنالتی $gap$ ($gap\ penalty$) گویند.

چند مثال:

۱:

$$\sigma(\mathrm{x}, \mathrm{y}) = \begin{cases} 2\ if\ x = y \\ 1\ if\ x \neq y \end{cases}$$

$$\sigma(\mathrm{X}, -) = \sigma(-, \mathrm{X}) = -1$$

۲:

امتیاز $m = match$

امتیاز $-s = mismatch$

امتیاز $-d = gap$

## ۲-۱ نحوه‌ی یافتن بهترین هم‌ترازی

ابتدا تعداد هم‌ترازی‌های ممکن را می‌شماریم. داریم:

$$C(n,m) = \sum_{k=0}^{k=\min(n,m)} \binom{n+m-k}{k, n-k, m-k}$$

اثبات. $k$ را تعداد مکان‌هایی که دو رشته با هم همردیف گردیده‌اند در نظر می‌گیریم. آن‌گاه در یک همردیفی، یک مکان یا متعلق به این $k$ مکان می‌باشد و یا اینکه $indel$ اتفاق افتاده است و مکان $deletion$ و $insertion$ نیز بایستی مشخص کردد.

اگر حتی به طور ساده نیز به مسئله نگاه کنیم، پیدا کردن همردیفی کار سختی می‌باشد، اما روش $dynamic\ programming$ برای آن وجود دارد که در جلسات بعدی به آن خواهیم پرداخت.