۱ ماتریس‌های امتیازدهی

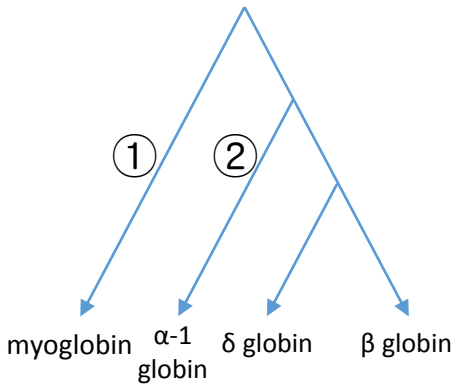
می‌دانیم که ماتریس امتیازدهی را می‌توان متناسب با مدل تکاملی بدست آورد اما در عمل ماتریس‌های مشخصی را برای امتیازدهی انتخاب می‌کنند که در این‌جا به آن‌ها پرداخته می‌شود. امتیازدهی در رشته‌های DNA ساده‌تر است و از همین رو، در این‌جا تنها به مبحث پروتئین‌ها پرداخته می‌شود. ماتریس‌های امتیازدهی برای سادگی با مقدار صحیح در نظر گرفته می‌شود. دو نوع ماتریس امتیازدهی برای پروتئین‌ها معرفی شده‌اند که در این‌جا به آن‌ها خواهیم پرداخت:

۱. (PAM) Point Accepted Mutation
۲. (BLOSUM) Block Substitution Matrix

۱-۱ ماتریس PAM (Dayhoff 1978)

ماتریس PAM بر پایه‌ی مفهومی به نام Accepted Point Mutation بنا نهاده شده است و بدان معناست که یک جهش در رشته‌ی آمینواسید ایجاد شده است و این جهش توسط محیط پذیرفته شده و در جامعه باقی مانده است.

برای تشخیص این جهش‌ها، Dayhoff et.al روش زیر را استفاده نمودند. آن‌ها ۳۴ خانواده پروتئینی را انتخاب کرده و از روی آن‌ها ۷۱ درخت phylogeny با استفاده از روش parsimony ترسیم نمودند. برای آن‌که تعداد جهش‌ها در یک مکان محدود باشد، پروتئین‌های نزدیک به یکدیگر را برای ساخت درخت استفاده نمودند (۸۵ درصد شباهت مورد استفاده قرار گرفته است). به عنوان مثال اگر به درخت globinها نگاه کنیم، تعداد جهش‌های پذیرفته شده را از روی آن می‌توانیم بدست آوریم. به عنوان مثال در تصویر زیر، ۴ مکان جهش وجود دارد و در مجموع ۸ جهش پذیرفته‌شده را مشاهده می‌کنیم.



$$C \rightarrow E\ A\ N\ L$$
$$1 \rightarrow A\ N\ K\ A$$
$$2 \rightarrow G\ L\ L\ V$$

برای وارد کردن زمان در مدل، Dayhoff et.al ماتریس PAM1 را معرفی نمودند که مبین زمانی است که یک درصد از آمینواسیدها دچار جهش می‌گردد.

به طور کلی می‌دانیم که امتیازات بر حسب $\lg\frac{P_{ij}}{P_iP_j}$ محاسبه می‌گردد. بنابراین با داشتن جهش‌های پذیرفته شده می‌توان نوشت:

نگارنده: سیدعباس حسینی
ایمیل: a_hosseini@ce.sharif.edu

که در آن $\hat{P}_{ij}$ تعداد جهش‌های پذیرفته‌شده $i \to j$ و $\hat{P}_i$ و $\hat{P}_j$ فرکانس مشاهده‌ی i و j در داده می‌باشد.

البته به دلیل آنکه هدف یافتن PAM1 می‌باشد، بهتر است در ابتدا ماتریس انتقال را تخمین بزنیم. با توجه به بازگشت‌پذیری زنجیره‌ی مارکوف می‌بایست داشته باشیم:

$$f(j)M(i,j) = f(i)M(j,i) \qquad *$$

به منظور تخمین $M(i,j)$ ابتدا کمیت‌های زیر را تعریف می‌کنیم:

$m(j)$ ← جهش‌پذیر بودن آمینواسید j

$f(j)$ ← فرکانس آمینواسید j

و در نتیجه $\sum_{j=1}^{20} f(j) = 1$. اگر N تعداد کل آمینواسیدها باشد، داریم

$$f(j) = \frac{n(j)}{N}$$

که $n(j)$ تعداد مشاهدات آمینواسید j در داده می‌باشد. هم‌چنین

$$m(j) = \frac{\sum_{i=1}^{20} A(i,j)}{n(j)}$$

که در آن $A(i,j)$ تعداد جهش‌های پذیرفته شده از i به j می‌باشد.

ماتریس $M(i,j)$ بیانگر احتمال جهش jامین آمینواسید به iامین آمینواسید می‌باشد. در نتیجه

$$M(i,j) = \lambda \frac{A(i,j)}{n(j)}$$

و بنابراین با توجه به این که $A(i,j) = A(j,i)$، رابطه‌ی * پذیرفته می‌شود. M محاسبه گردیده و برای زمان متناسب با ۱ جهش در صد آمینواسید انتخاب می‌گردد و در نتیجه می‌توان از روی آن ماتریس امتیازدهی $PAM1$ را نیز محاسبه نمود.

۱-۲ PAMn:

به دلیل مارکوف بودن فرایند، برای n جهش در ۱۰۰ آمینواسید داریم:

$$M_n = M_1^n$$

و ماتریس امتیازدهی به صورت

$$PAM_n(i,j) = \log \frac{f(j)M_n(i,j)}{f(i)f(j)}$$

محاسبه می‌گردد.

۱-۳ BLOSUM (Henikoff & Henikoff 19992)

برای یافتن ماتریس PAM، پروتئین‌های نزدیک به یکدیگر انتخاب گردیدند. در روش BLOSUM تمام پروتئین‌های یک خانواده با یکدیگر در نظر گرفته شده و نواحی که در خانواده باقی مانده است، برای تعیین ماتریس مورد استفاده قرار گرفته است. هم‌چنین نواحی که بسیار نزدیک به یکدیگر می‌باشند نیز متوسط‌گیری می‌گردد تا از ایجاد بایاس جلوگیری شود.

نکته: با گروه‌بندی انجام‌گرفته می‌توان فرکانس‌ها را محاسبه نمود و تخمین ماتریس امتیازدهی را بدست آورد. به عنوان مثال در BLOSUM62 تمام بلوک‌های پروتئینی که حداقل 62% یکسانی داشته باشند را در یک گروه قرار داده و جهش را مجاسبه می‌نماییم.

نکته: برای یافتن BLOSUM n بر خلاف PAM که از یک ماتریس استفاده می‌شد، مستقیما از داده استفاده می‌شود.

نکته: به طور کلی می‌توان گفت که بهتر است از ماتریس PAM برای هم‌ردیفی کلی و برای BLOSUM برای هم‌ردیفی محلی استفاده کرد.

نحوه‌ی استفاده: اگر هدف هم‌ردیفی کلی است و رشته‌های مشابه را مطالعه می‌کنیم، PAM 250 مناسب است و اگر چیزی نمی‌دانیم PAM 120 بهتر است. برای هم‌ردیفی محلی به طور عموم از BLOSUM 62 استفاده می‌گردد.

۲ ارزیابی هم‌ردیفی‌های محلی

(چگونه آستانه را انتخاب نماییم؟)

فرض کنید که دو رشته‌ی X^n و Y^m را که به طور تصادفی تولید شده‌اند در اختیار دارید. هم‌چنین فرض کنید که ماتریس امتیازدهی استفاده شده برای تشخیص Homolog را نیز در اختیار داریم. یعنی

$$\sigma_{ij} = \log \frac{P(x_i, y_j)}{P(x_i, y_j)}$$

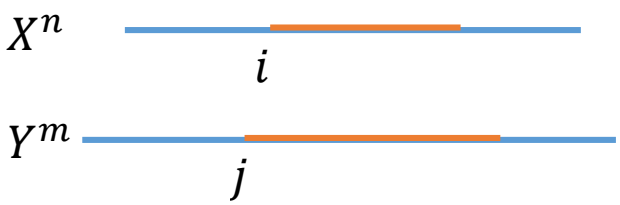
موجود است. حال اگر دو زیررشته از رشته‌های X^n و Y^m را بدون gap هم‌ردیف نماییم، می‌توانیم امتیاز هم‌ردیفی را محاسبه نماییم. اگر این امتیاز را برای S قرار داده و تعداد هم‌ردیفی‌های بدون gap برای X^n و Y^m را که امتیاز حداقلی S را دارا می‌باشند، با N_s نشان دهیم آنگاه N_s تقریبا دارای توزیع پواسون بوده و متوسط آن برابر با

$$E(N_S) \approx kmn\, e^{-\lambda S}\ [Karlin\ \&\ Altschvl]$$

خواهد بود که در آن λ و k از روی ماتریس امتیاز بدست می‌آید.

برای این که شهودی از قضیه داشته باشیم، می‌بینیم که برای هر نقطه آغازین i و j، احتمال داشتن امتیازی بزرگ‌تر از S تقریبا برابر $e^{-\lambda S}$ بوده و در نتیجه تعداد متوسط $E(N_S)$ خواهد بود. حال اگر بخواهیم احتمال آن‌که یک هم‌ردیفی بدون gap با امتیاز حداقل S بیابیم، آنگاه

$$\begin{aligned} &P(S(X^n, Y^m) > S) \approx P\{N_S \geq 1\} \\ &= 1 - P\{N_S = 0\} \\ &= 1 - e^{-kmn\, e^{-\lambda S}} \\ &= e^{-E} \end{aligned}$$



به طور کلی با نرمالایز کردن امتیاز S به

$$bit - score \rightarrow S' = \frac{\lambda S - \ln k}{\ln 2}$$

داریم

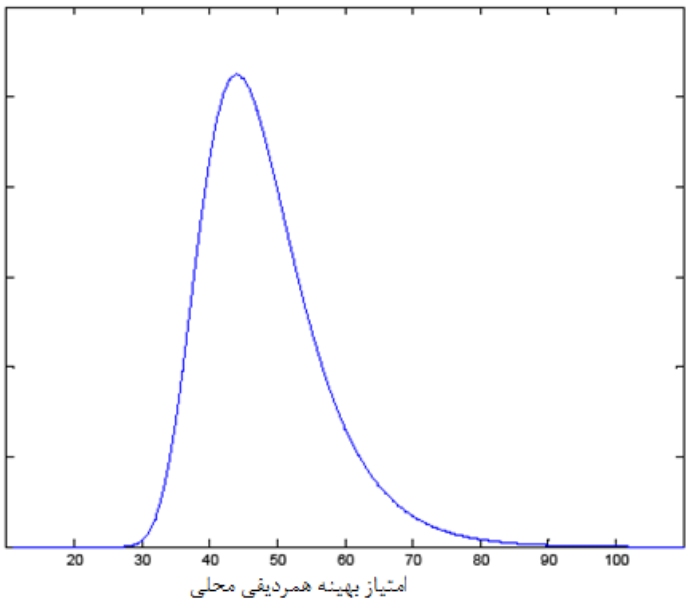
$$.E = mn\, e^{-S'}$$

اگر به توزیع امتیاز نگاه کنیم، داریم:

$$F(S) = P(S(X^n, Y^m) < S) = e^{-kmn\, e^{-\lambda S}}$$

که توزیع Gumbel نام دارد. در حالتی که gap را نیز در نظر بگیریم، تئوری قوی وجود ندارد که توزیع امتیازات از Gumbel پیروی نماید اما شبیه‌سازی زمانی انجام‌شده بر روی رشته‌های تصادفی به صورت تجربی داده‌ها با توزیع Gumbel هم‌خوانی دارد.

مثال: ۱۰۰۰۰ جفت تصادفی از رشته‌هایی پروتئینی با طول ۱۰۰۰ تولید می‌نماییم. اگر از ماتریس BLOSUM62 استفاده گردد، آنگاه با دادن امتیاز منفی $k - 11$ برای gapها (یعنی امتیاز $11-$ برای آغاز gap و $1-$ برای تعداد gapها) به توزیع



تصویر ۱

می‌رسیم که با پارامترهای $\lambda = 0.27$ و $k \approx 0.04$، با توزیع Gumbel مشابهت می‌یابد.

مثال: فرض می‌نماییم که یک query با طول ۲۰ را با پایگاه داده‌ای به طول ۳۰ هم‌ردیفی محلی نموده‌ایم و هم‌ردیفی

$$query \to actggtccat$$
$$subject \to actgat--at$$

حاصل گردیده است که در آن ۱جابجایی، ۷ تطابق و ۲ gap وجود دارد. اگر امتیازدهی به صورت

$$S = aI + bX - cO - dG$$

باشد که I تعداد تطابق، X تعداد جابجایی، O تعداد opening gap و G تعداد extension gap باشد و ضرایب به صورت $a = 3, b = -3, c = 5, d = 2$ در نظر گرفته شود، $S = 5$ خواهد شد. با در نظر گرفتن $k = 0.711$ و $\lambda = 1.37$ خواهیم داشت

$$S' = \frac{\lambda S - \ln k}{\ln 2} = 12.82$$

و در نتیجه $E = mn2^{S'} = 0.0016$ بدست می‌آید. هرچه میزان E کوچکتر باشد بهتر است، زیرا نشان می‌دهد که در دو رشته تصادفی امکان وقوع دیدن امتیازی بزرگ‌تر از S بسیار کم است.