1. جهش Mutation

همان‌گونه که می دانیم، تغییراتی که در **DNA** رخ می‌دهند باعث تغییر در ژنوم و سرآخر **Phenotype** می‌گردند و هر جهش **L** را به یک رشتهٔ دیگر تبدیل می‌نماید. در این قسمت ابتدا جهش را مدل کرده و اینکه جهش چه بلایی سر رشته در می‌آورد را در نظر نمی‌گیریم. در حقیقت جهش‌ها باعث تغییر در **fitness** گردیده و مدل **WF** و یا **Coalescent** را به هم می‌ریزند. به هر حال در این قسمت فرض ما بر این است که جهش‌ها **Neutral** بوده و تأثیری در **fitness** ندارند. با این فرض‌ها می‌توان جهش‌ها را یک فرآیند مستقل نسبت به **Coalescent** در نظر گرفت.

۱.۱ تعریف مدل

در مدل **WF** با جهش، جهش با احتمال **u** بین نسل‌ها اتفاق می‌افتد.

بدین ترتیب تعداد جهش‌ها از یک نسل به نسل دیگر با توزیع دو جمله‌ای مدل می‌گردند:

$$P(K=\#Mutation \text{ in } i_{th} \text{ generation})=\binom{2N}{K}x^K(1-u)^{2N-K}$$

قضیه:

در مدل **Coalescent** ، جهش‌ها با توزیع پواسون بر روی شاخه‌های درخت به وقوع می‌پیوندند و نرخ جهش برابر است با

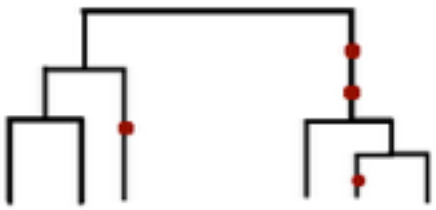
$$\frac{\theta}{2}=2Nx$$

اثبات:

اگر یک شاخه را در مدل **Coalescent** در نظر بگیریم، زمانی که نیاز است منتظر بمانیم تا اولین جهش رخ دهد برابر است با:

$$P(T>t)=(1-u)^{t2N}=(1-\frac{\theta}{4N})^{t2N}\to e^{-\frac{\theta t}{2}}$$

و در نتیجه **T** توزیع نمایی با پارامتر $\frac{\theta}{2}=2Nu$ دارد. بنابراین یک فرآیند پواسون را برای یک شاخه خواهیم داشت. از طرف دیگر جهش‌ها در شاخه‌های دیگر هم مستقل هستند و آن‌ها نیز دارای فرآیند پواسون هستند.



در نتیجه تعداد جهش‌ها بر روی یک شاخه درخت به طول **L** دارای توزیع پواسون با متوسط $\frac{\theta L}{2}$ خواهد بود. در نتیجه الگوریتم زیر را برای مدل کردن جهش‌ها می‌توان استفاده نمود.

نگارنده: افسون افضل
ایمیل: afs.afzal@gmail.com

۱.۲. الگوریتم ساختن درخت Coalescent با جهش

۱- درخت **Coalescent** را شبیه‌سازی کن.

۲- برای هر شاخه با طول **L** ، تعداد جهش‌ها را از روی توزیع پواسون با متوسط $\frac{\theta L}{2}$ بدست آور.

۳- بر روی هر شاخه، زمان‌های وقوع جهش به طور یکنواخت بر روی پخش شده‌اند.

قضیه:

در درخت **Coalescent** با جهش، اتفاقات (چه از جنس **Coalescent** و چه جنس جهش) با نرخ $\frac{k(k-1+\theta)}{2}$ وقوع می‌پذیرند که **k** تعداد **lineage**ها است. زمانی که یک اتفاق رخ داد آنگاه با احتمال $\frac{\theta}{k-1+\theta}$ جهش بوده و با احتمال $\frac{k-1}{k-1+\theta}$ **Coalescent** است.

اثبات:

در حقیقت دو توزیع نمایی مستقل **x**, **y** بانرخ‌های λ_1 و λ_2 را در نظر مگیریم. آنگاه $\min(x,y)$ دارای توزیع نمایی با نرخ $\lambda_1+\lambda_2$ خواهد بود.

$$P(\min(x,y)<t)=P(x<t)+P(x>t)P(y<t)=1-e^{-(\lambda_1+\lambda_2)t}$$

در نتیجه زمان انتظار برای وقوع یک حادثه دارای توزیع نمایی با پارامتر $\binom{k}{2}+\frac{\theta k}{2}=\frac{k(k-1+\theta)}{2}$ است. از طرف دیگر برای آنکه ببینیم کدام زودتر اتفاق می‌افتد کافیست که بدانیم:

$$P(x<y)=\int_0^\infty f_x(x)[1-F_2(x)]dx=\int_0^\infty \lambda_1 e^{-\lambda_1 x}e^{-\lambda_2 x}\,dx=\frac{\lambda_1}{\lambda_1+\lambda_2}$$

و قضیه فوق اثبات می‌شود.

۱.۳ شبیه‌سازی درخت Coalescent با جهش (روش دوم)

۱- با **n=k** که **n** تعداد نمونه‌ها است آغاز کن.

۲- یک زمان نمایی با پارامتر $\frac{k(k-1+\theta)}{2}$ ایجاد کن.

۳- با احتمال $\frac{(k-1)}{(k-1+\theta)}$ یک **Coalescent** است و در غیر این صورت جهش است.

۴- اگر **Coalescent** اتفاق افتاده بود، دو تا از خطوط را تصادفی انتخاب کن و آن‌ها را به هم وصل کن.

۵- اگر جهش اتفاق افتاده بود یک خط را به تصادف انتخاب کن و جهش را در آن ایجاد نما.

۶- اگر $k>1$ به گام ۲ بازگرد.

قضیه (Watterson۷۵-Tavare۸۴)

اگر S_n را تعداد جهش‌های یک درخت با n ژن قرار دهیم آنگاه:

$$P(S_n=S)=\frac{n-1}{\theta}\sum_{i=0}^{n-1}(-1)^{i-1}\binom{n-2}{i-1}\left(\frac{\theta}{i+\theta}\right)^{S+1}$$

اثبات:

اگر طول تمام شاخه‌ها را برابر با T_{total} در نظر بگیریم، آنگاه با توجه به T_{total} تعداد جهش‌ها دارای توزیع پواسون خواهد بود.

$$P(S_n=s\,|\,T_{total}=t)=\frac{(\theta\frac{t}{2})^s}{s!}e^{-\frac{\theta t}{2}}$$

در نتیجه

$$P(S_n=s)=\int_0^{\infty}\frac{(\frac{\theta t}{2})^s}{s!}e^{-\frac{\theta t}{2}}P(T_{total}=t)dt$$

از طرفی

$$P_{total}(t)=\sum_{i=2}^{n}\frac{i-1}{2}e^{-\frac{i-1}{2}t}\prod_{j=2,j\neq i}^{n}\frac{j-1}{j-i}*$$

با قرار دادن رابطهٔ فوق جواب مسئله بدست می‌آید.

برای آنکه * را اثبات کنیم کافیست بدانیم که اگر $x_1,x_2,...,x_n$ دارای توزیع نمایی، با پارامترهای $\lambda_1,\lambda_2,...,\lambda_n$ باشند، آنگاه جمع آن‌ها دارای توزیع زیر است:

$$P_{\sum x_i}(x)=\sum_{i=1}^{n}\lambda_i e^{-\lambda_i x}\prod_{j=1,j\neq i}^{n}\frac{\lambda_j}{\lambda_j-\lambda_i}$$

بنابراین در درخت Coalescent تعداد شاخه‌ها و تعداد وقوع Coalesce طول کل شاخه‌ها را نشان می‌دهد.

$$T_{total}=\sum_{i=2}^{n}T_i$$

توزیع iT_i نمایی با پارامتر $\frac{i-1}{2}$ بوده و با استفاده از قضیه فوق توزیع بدست می‌آید.

می‌توان از رابطهٔ بازگشتی زیر نیز استفاده نمود:

$$P(S_n = s) = \frac{n-1}{n-1+\theta} P(S_{n-1} = s) + \frac{\theta}{n-1+\theta} P(S_n = s-1)$$

در حقیقت در هر مرحلهٔ اتفاق، یا اتفاق جهش است که از **s** یک واحد کم می‌گردد و یا **Coalesce** بوده که از تعداد **lineage** ها (**n**) یک واحد کم می‌گردد. با توجه به آنکه $P(S_1 = 0) = 1$ رابطه بازگشتی فوق را می‌توان محاسبه نمود.

متوسط و واریانس S_n برابر خواهد بود با:

$$E(S_n) = \theta \sum_{i=1}^{n-1} \frac{1}{i} = \theta h_n \; *$$

$$Var(S_n) = \theta \sum_{i=1}^{n-1} \frac{1}{i} + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2} = \theta h_n + \theta^2 g_n$$

از رابطه * می‌توان استفاده نمود و تخمین **Watterson** را برای θ به صورت زیر نوشت:

$$\theta_w = \hat{\theta} = \frac{S_n}{h_n}$$

این تخمین‌گر دارای شرایط زیر است:

$$E[\theta_W] = \theta, Var(\theta_W) = \frac{\theta}{h_n} + \theta^2 \frac{g_n}{h_n^{\,2}}$$

مثال:

در سال ۱۹۹۵ رشته‌های مربوط به ژن **ZFY** را در ۳۸ فرد توالی‌یابی نمودند و هیچ‌گونه جهشی در آن پیدا نکردند. می‌خواهیم از این داده استفاده کرده و زمان **T** یعنی زمان تا جد مشترک را با توجه به داده تخمین بزنیم.

← می‌دانیم که $T = \sum_{i=2}^{38} T_i$ و همچنین $S_i = 0$. توزیع T_iها نمایی با پارامتر $\binom{i}{2}$ است. همچنین:

$$P(S_i = 0 \mid T_i = t) = e^{-\frac{t\theta i}{2}}$$

در نتیجه

$$P(T_i = t \mid S_i = 0) = \frac{P(S_i = 0 \mid T_i = t) P(T_i = t)}{P(S_i = 0)} \propto e^{-\frac{ti(\theta+i-1)}{2}}$$

پس توزیع T_i به شرط $S_i = 0$ نمایی است. در نتیجه

$$E(T|S=0) = \sum_{i=2}^{n} \frac{2}{i(\theta+i-1)}$$

اگر فرض کنیم که $2N_{eff} = 5000, u = 2\times10^{-5}$ خواهیم داشت:

$$\theta = 2N_{eff} = 0.2$$
$$E(T|S=0) = 1.72$$

اگر فرض کنیم که هر نسلی ۲۰ سال عمر می‌کند چیزی معادل ۱۷۲۰۰۰ سال تا جد مشترک فاصله خواهد بود.