估算群体中的 indel 附近
单核苷酸突变率相对增加的模型

假设 indel 的发生率 μ_{i} 很小, 并且是中性的. 对于 indel 和 nonindel 等位基因, 令 μ 和 μ_{het} 分别代表纯合子和杂合子中的中性核苷酸突变, 并且令 $f = \mu_{\text{het}}/\mu$. 指定 N_{i} 和 N_{ni} 分别代表 indel 和 nonindel 附近自 MRCA 之后的中性突变数. 我们要寻求期望值 $E(N_{\text{i}})$ 和 $E(N_{\text{ni}})$.

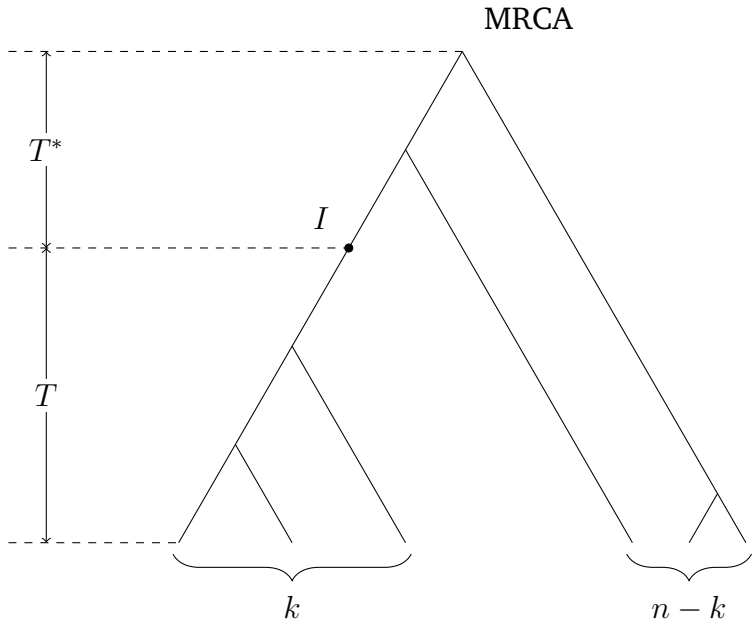

Figure 1: n 个基因的样本中 k 个基因含有中性 indel I 的中性系谱. 回溯到 I 的时间为 T, T^* 代表回溯到样本共同祖先 MRCA 的额外的时间. I 处的世系数为 ν.

在图 1 的系谱中, n 个基因的样本中 k 个携带 indel, 由 $k \geq 1$ 且 $n-k \geq 1$, 可知 $1 \leq k \leq n-1$; I 处的世系数为 ν, 由 $\nu \geq 2$ 且 $n-k \geq \nu-1$, 可知 $2 \leq \nu \leq n-k+1$. 以 $2N_{\text{e}}$ 个世代为单位衡量所有时间, 其中 N_{e} 代表有效群体大小. 回溯到 I 的时间是 T, T^* 代表回溯到样本共同祖先 MRCA 的额外的时间. 忽略 k 和 n 的相关性.

根据标准的中性突变模型 (Kimura 1983;Watterson 1975)有:

$$E(S) = \mu E(T_{\text{tot}}). \tag{1}$$

这个模型的基本假设是:

- 每个子代相对于亲代的突变个数服从泊松分布, 平均值为 μ;
- μ 为常数, 与群体大小、基因型和时间无关;
- 无选择和重组.

对于一个 indel, T_{het} 和 T_{hom} 分别代表 indel 在杂合子和纯合子中经历的时间. 可知:

$$T = T_{\text{het}} + T_{\text{hom}}, \tag{2}$$

根据方程 1 和 2 可以得到 $E(N_{\text{i}})$:

$$\begin{aligned} E(N_{\text{i}}) &= 2N_{\text{e}}[\mu_{\text{het}} E(T_{\text{het}}) + \mu E(T_{\text{hom}}) + \mu E(T^*)] \\ &= 2N_{\text{e}}\mu[\frac{\mu_{\text{het}}}{\mu} E(T_{\text{het}}) + E(T_{\text{hom}}) + E(T^*)] \\ &= \frac{\theta}{2}[f E(T_{\text{het}}) + E(T_{\text{hom}}) + E(T^*)] \\ &= \frac{\theta}{2}[f E(T_{\text{het}}) + f E(T_{\text{hom}}) - f E(T_{\text{hom}}) + E(T_{\text{hom}}) + E(T^*)] \\ &= \frac{\theta}{2}[f E(T) + E(T^*) - (f-1) E(T_{\text{hom}})], \end{aligned} \tag{3}$$

其中 $\theta = 4N_{\text{e}}\mu$.

在过去的 t 时刻, indel 的频率是 $X(t)$, 则:

- indel 在杂合子中的条件概率为 $1 - X(t)$;
- indel 在纯合子中的条件概率为 $X(t)$;
- nonindel 在杂合子中的条件概率为 $X(t)$;
- nonindel 在纯合子中的条件概率为 $1 - X(t)$.

在 t 时刻 indel 纯合性的指示函数为:

$$\chi_{\text{hom}}(t) = \begin{cases} 1 & \text{如果 indel 在时刻 } t \text{ 处于纯合子,} \\ 0 & \text{如果 indel 在时刻 } t \text{ 处于杂合子.} \end{cases} \tag{4}$$

因此,

$$T_{\text{hom}} = \int_0^T \chi_{\text{hom}}(t)dt. \tag{5}$$

这导出了条件期望

$$\begin{aligned} E[T_{\text{hom}}|X(0)=x] &= E\left\{E\left[\int_0^T \chi_{\text{hom}}(t)dt \;\middle|\; T, X(0)=x\right]\right\} \\ &= E\left\{\int_0^T E[\chi_{\text{hom}}(t)]dt \;\middle|\; X(0)=x\right\} \\ &= E\left\{\int_0^T X(t)dt \;\middle|\; X(0)=x\right\}. \end{aligned} \tag{6}$$

类似地, 可得

$$E[T_{\text{het}}|X(0)=x] = E\left\{\int_0^T [1-X(t)]dt \;\middle|\; X(0)=x\right\}. \tag{7}$$

交换 $E[T_{\text{het}}]$ 和 $E[T_{\text{hom}}]$, 用上一段的推导方法可得

$$\begin{aligned} E(N_{\text{ni}}) &= \frac{\theta}{2}[fE(T_{\text{hom}}) + E(T_{\text{het}}) + E(T^*)] \\ &= \frac{\theta}{2}[E(T) + E(T^*) + (f-1)E(T_{\text{hom}})]. \end{aligned} \tag{8}$$

这样, 我们必须计算方程 3 和 8 里的三个平均时间. 自此以后, 假设 $\theta_{\text{i}} = 4N_{\text{e}}\mu_{\text{i}} \to 0$, 并忽略 indel 到 nonindel 的回复突变.

为了计算方程 6, 首先要注意 indel 在群体中还没有固定 ($0 \le k \le n$), 可用 Maruyama and Kimura (1975) 的方程 16 来计算一个 indel 固定前的平均时间:

$$\lim_{p\to 0}\frac{F_x^{(1)}(p)}{\Phi(p,x)} = \frac{4N_{\text{e}}}{1-4N_{\text{e}}\nu}\left\{\int_0^x \frac{1-(1-\xi)^{1-4N_{\text{e}}\nu}}{\xi}d\xi + \frac{1-(1-x)^{1-4N_{\text{e}}\nu}}{(1-x)^{1-4N_{\text{e}}\nu}}\int_x^1 \frac{(1-\xi)^{1-4N_{\text{e}}\nu}}{\xi}d\xi\right\}.$$

根据我们的需要对这个方程做修改: (i) 除以 $2N_{\text{e}}$ 来衡量时间; (ii) 突变率趋近于 0, 即 $4N_{\text{e}}\nu$ 趋近于 0; (iii) 在两个被积函数中都插入一个因子 ξ, 即同乘以 ξ. 可以得到:

$$\begin{aligned} E[T_{\text{hom}}|X(0)=x] &= 2\int_0^x \xi d\xi + \frac{2x}{1-x}\int_x^1 (1-\xi)d\xi \\ &= x^2 + \frac{2x}{1-x}\left(\frac{1}{2}(x-1)^2\right) \\ &= x. \end{aligned} \tag{9}$$

为了从方程 9 中推导出 $E[T_{\text{hom}}]$, 需要得到在 n 个基因的样本中观察到 k 个 indels 的概率密度 $X(0)$. 根据贝叶斯公式,群体频谱为 θ/x (Kimura 1969, 1971; Ewens 2004),

样本频谱为 θ_{i}/x (Watterson 1975; Ewens 2004), 可以得条件密度为

$$\begin{aligned}\phi(x) &= \binom{n}{k} x^k (1-x)^{n-k} \frac{\theta_{\mathrm{i}}}{x} \Big/ \frac{\theta_{\mathrm{i}}}{k} \\ &= k\binom{n}{k} x^{k-1}(1-x)^{n-k}.\end{aligned} \tag{10}$$

结合方程 9 和 10 得到

$$a \equiv E(T_{\mathrm{hom}}) = \frac{k}{n+1}. \tag{11}$$

从 Wiuf and Donnelly (1999) 得知:

$$b \equiv E(T) = \frac{2k}{n-1} - \frac{2}{n} + 2\binom{n-1}{k}^{-1} \sum_{j=2}^{n-k+1} \frac{1}{j}\binom{n-j-1}{k-1}. \tag{12}$$

对于 $E(T^*)$ 的计算中需要用到 I 时刻的世系数目 ν 。用 T_{j1} 代表 j 个世系到 MRCA 的时间. 它的平均值为 (Kingman 1982):

$$E(T_{j1}) = 2\left(1 - \frac{1}{j}\right). \tag{13}$$

由此可知:

$$E(T^*) = E[E(T_{\nu 1}|\nu)] = 2E\left(1 - \frac{1}{\nu}\right). \tag{14}$$

从 Wiuf and Donnelly (1999) 的方程 18 可以知道 ($2 \le j \le n-k+1$):

$$P(\nu = j) = \binom{n-j}{k-1}\binom{n-1}{k}^{-1}. \tag{15}$$

将方程 15 代入方程 14 得到:

$$c \equiv E(T^*) = 2\binom{n-1}{k}^{-1} \sum_{j=2}^{n-k+1} \left(\frac{j-1}{j}\right)\binom{n-j}{k-1}. \tag{16}$$

方程 3 除以方程 3, 再代入 (11), (12) 和 (16), 化简得到

$$r \equiv \frac{E(N_{\mathrm{i}})}{E(N_{\mathrm{ni}})} = \frac{fb + c - (f-1)a}{b + c + (f-1)a}. \tag{17}$$

我们已知 a, b 和 c, 并可以用 $N_{\mathrm{i}}/N_{\mathrm{ni}}$ 来估计 r, 以 f 为未知数解方程 17 得

$$f = \frac{a + c - r(b + c - a)}{a(r+1) - b}. \tag{18}$$

可知函数 $f(r)$ 的图像为双曲线,

- 当 $r = 1$ 时, $f = 1$;
- 当 $r \to (b-a)/a$ 时, $f \to \infty$;
- 当 $r < (a+c)/(b+c-a)$ 时, $f < 0$.

根据参数的实际意义, 要求 $f > 0$, r 的取值范围为 $(\frac{a+c}{b+c-a}, \frac{b-a}{a})$.

将方程 17 改写为

$$r = \frac{b-a}{a} - \frac{(b-2a)(b+c)}{a(b+c-a+fa)}. \tag{19}$$

为了让 r 有意义, 要求满足 $b > 2a$.

用方程 9 类似的推导方法, 这次在 Maruyama and Kimura (1975) 的方程 16 中插入因子 $1-\xi$, 可得

$$\begin{aligned} E[T_{\text{het}}|X(0)=x] &= 2\int_0^x (1-\xi)d\xi + \frac{2x}{1-x}\int_x^1 \frac{(1-\xi)^2}{\xi}d\xi \\ &= -x\left(1 + \frac{2\ln x}{1-x}\right) \\ &> -x + \frac{2x(1-x)}{1-x} = x. \end{aligned} \tag{20}$$

比较方程 20 与 9, 可知

$$E[T_{\text{het}}|X(0)=x] > E[T_{\text{hom}}|X(0)=x]. \tag{21}$$

从方程 1 与 21, 得到

$$E[T|X(0)=x] > 2E[T_{\text{hom}}|X(0)=x]. \tag{22}$$

现在方程 11, 12 和 22 意味着 $b > 2a$.

前述结果表明, 从方程 18 估计 f 时, 将期望一个较大的置信区间.

上面是从每个 k 来得到 f 的估计值. 因为分子生物学参数 f 的真实的值必须与样本参数 k 无关, 下面是从完整的数据集中得到 f 的估计值的方法. 将方程 3 和 8 转换为依赖于 k 的形式:

$$\begin{aligned} E(N_{\text{i}}^{(k)}) &= \frac{\theta}{2}[fE(T^{(k)}) + E(T^{*(k)}) - (f-1)E(T_{\text{hom}}^{(k)})] \\ &= \frac{\theta}{2}[fb^{(k)} + c^{(k)} - (f-1)a^{(k)}], \\ E(N_{\text{ni}}^{(k)}) &= \frac{\theta}{2}[b^{(k)} + c^{(k)} + (f-1)a^{(k)}]. \end{aligned} \tag{23}$$

引入方程 11, 12 和 16, 并重新定义:

$$
\begin{aligned}
E(N_{\mathrm{i}}) &= \sum_{k=1}^{n-1} E(N_{\mathrm{i}}^{(k)}), \\
E(N_{\mathrm{ni}}) &= \sum_{k=1}^{n-1} E(N_{\mathrm{ni}}^{(k)}), \\
a &= \sum_{k=1}^{n-1} a^{(k)} = \frac{n(n-1)}{2(n+1)}, \\
b &= \sum_{k=1}^{n-1} b^{(k)} = n - 2 + \frac{2}{n} + 2\sum_{j=2}^{n} \frac{1}{j} \sum_{k=1}^{n-j+1} \binom{n-1}{k}^{-1} \binom{n-j-1}{k-1}, \\
c &= \sum_{k=1}^{n-1} c^{(k)} = 2\sum_{j=2}^{n} \frac{j-1}{j} \sum_{k=1}^{n-j+1} \binom{n-1}{k}^{-1} \binom{n-j}{k-1},
\end{aligned}
\tag{24}
$$

方程 23 对 k 累加, 得到方程 24 中的重新定义值 a, b, c. 方程 17 和 18 可得完整的数据集中 f 的估计值.

专用词与符号

缩写

DNA Deoxyribonucleic acid, 脱氧核糖核酸

MRCA Most Recent Common Ancestor, 最近共同祖先

Indel Insertion/Deletion, 插入/缺失

符号

μ The mean number of mutations, 突变数的平均值

t The number of generations since the MRCA of two sampled homologous sequences, 从 MRCA 到两个抽样的同源序列的世代数

S The number of mutations that have occurred in the descent to the two descendent sequences, 向下到这两个后代的序列过程中产生的突变数

T_{tot} The sum of the lengths of the branches of the genealogy of a sample, 一个样本的系谱中所有分支长度的总和

函数

$E()$ Expected value, 期望值

Var() Variance of expected value, 期望值的方差

名词

Alleles 等位基因

Bayes' formula 贝叶斯公式

Conditional expectation 条件期望

Confidence interval 置信区间

Distritution 分布

Equation 方程

Frequency spectrum 频谱

Heterozygote 杂合子

Homozygote 纯合子

Indicator function 指示函数

Lineage 世系

Mutations 突变

Probability density 概率密度

Sample 样本

参考文献

Ewens, Warren John. 2004. *Mathematical Population Genetics: Theoretical Introduction*. Springer.

Kimura, M. 1969. “The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population due to Steady Flux of Mutations”. *Genetics* 61 (4): 893–903.

— . 1971. “Theoretical Foundation of Population Genetics at the Molecular Level”. *Theoretical Population Biology* 2 (2): 174–208.

Kingman, J. F. C. 1982. “On the Genealogy of Large Populations”. *Journal of Applied Probability* 19:27.

Maruyama, T., and M. Kimura. 1975. “Moments for Sum of an Arbitrary Function of Gene Frequency along a Stochastic Path of Gene Frequency Change”. *Proceedings of the National Academy of Sciences of the United States of America* 72 (4): 1602–1604.

Watterson, G. A. 1975. “On the Number of Segregating Sites in Genetical Models without Recombination”. *Theoretical Population Biology* 7 (2): 256–276.

Wiuf, C., and P. Donnelly. 1999. “Conditional Genealogies and the Age of a Neutral Mutant”. *Theoretical Population Biology* 56 (2): 183–201.