模型的建立和下一步的目标

数 33 赵丰会议记录

January 12, 2017

1 模型建立

假设在两组实验条件下分别获得 RNA 序列的统计信号 $Y_1(n)$ 和 $Y_2(n)$. 其中 n 是正整数, 表示位点的序号 $Y_1(n) \geq 0, 1 \leq n \leq N$,N 表示序列长度。对给定的 RNA 序列我们假设其在不同的实验条件下为 Markov 链，分别记为 $X_1(n), X_2(n), X_i(n), i = 1, 2$ 只能取 0 和 1 两个取值 (可以设 0 表示单链，1 表示双链)。每条链的初始状态为 $X_i(1)$，是非随机的。每条链的转移概率矩阵为 $P_i \square P_i$ 具有的特征是对角元的数特别大（比如为 0.99), 表示该链上后一个位点的状态具有很大的概率和前一个位点的状态保持一致。参照 BUMHMM Model 我们进一步假设当 $X_i(n) = 0$ 时，$Y_i(n)$ 服从 $Beta(\alpha, \beta)$ 分布，当 $X_i(n) = 1$ 时，$Y_i(n)$ 服从均匀分布。这样即建立原问题的数学模型。

2 求解思路

参考 $Data_Processing.pdf$, 为降低单位点实验重复次数较少带来的误差，我们考虑 r(比如 r=3) 个位点联合比较，这时做假设检验的统计量具有如下基本的形式：

$$\sum_{i=1}^{r} \alpha_i \frac{\bar{z_i} - \bar{z_i'}}{\sqrt{S_{z_i}^2/N1 + S_{z_i}'^2/N1}} \tag{1}$$

其中 $N1$ 为实验重复次数，α_i 是线性组合系数，一般可全取 1。以下有两种主要思路：

(1) 假设相邻 r 个位点的相关性很强（要么“全相同”，要么全不同，这里“全相同”指的是两条链的对应 r 个位点同时为单链或双链）由于同一个位点 (忽略端点）可以参与 r 组假设检验，因此可以得到 r 个是否接受原假设的判决结果（设接受原假设为 0, 拒绝原假设为 1），根据 r 个值中 0 占多数还是少数给出最终某位点在不同实验条件下单双链是否相同的判决，同时根据 0,1 的比例报告置信度。

(2) 假设相邻 r 个位点的有一定的相关性（例如 r=3 时不会出现相同-不同-相同这种组合），则需要构造一个多个结果的假设检验，由相关性的假设可以认为假设检验结果的个数小于 2^r, 为判断最终结果属于哪一类，需要有多个统计量参与，构造多个统计量的方法可参考 (1) 式，只是取 α_i 为不同的值（比如相互正交)。在这种情况下，需要先把长度为 N 的链等分成不相交的若干份，每份长度为 r，对每份分别进行处理。

3 下一步的打算

1. 首先用 R 语言尝试正向仿真建模。
2. 其次用上述方法尝试给出判决结果，并与事前数据进行比较。
3. 如果上述假设检验方法较合理，尝试对有训练集的短链 RNA 进行模型的参数匹配。
4. 由于上述模型比较复杂，完整的理论分析比较困难，但会尝试给出某些理论分析的结果。

4 参考文献

References

[1] 生物信息学