# 模型的分析和仿真的结果

数 33 赵丰

January 14, 2017

## 1 模型分析

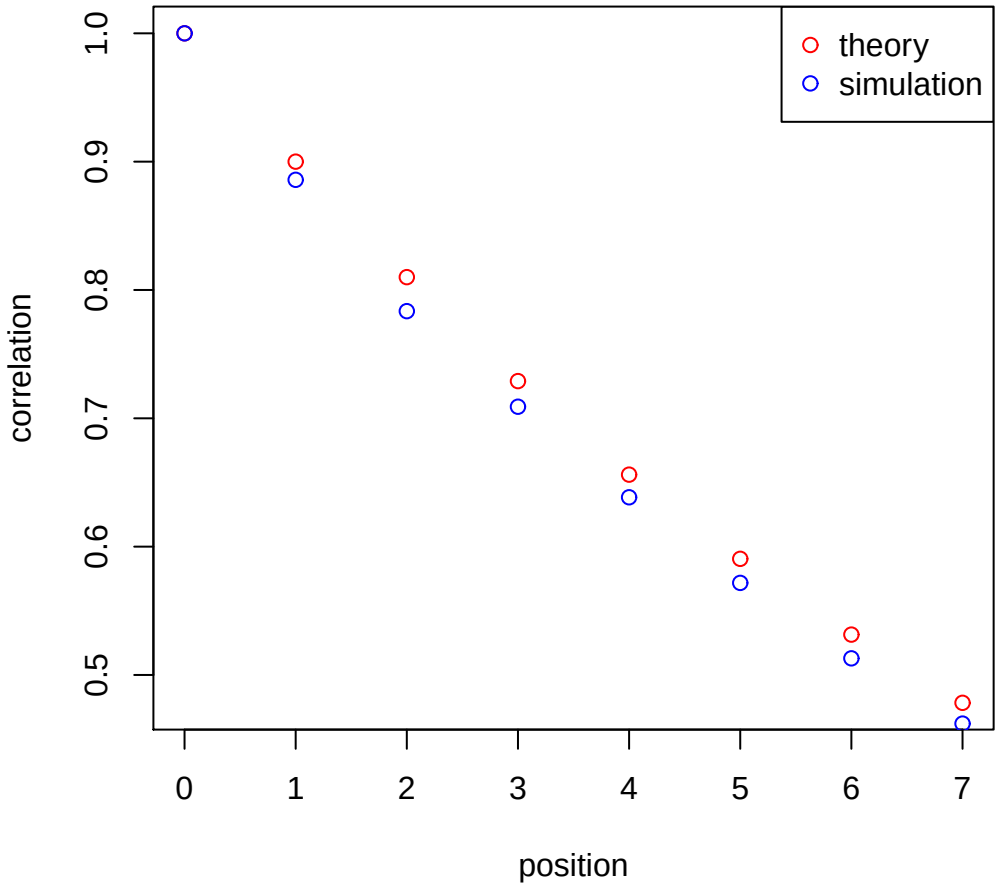
虽然 RNA 序列由于自身碱基配对而使得序列的相关性分析更复杂，但对于其中一小段较短的片段，可以近似认为其具有如下的单步转移概率矩阵:

$$P = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}$$

其中 p 接近 1. 由转移矩阵 P 可以直接求出 $R(1) = \frac{p}{2}$, 其相关系数为 $\rho(1) = \frac{R(1)-E(X_n)E(X_{n-1})}{\sigma^2_{X_n}} = 2p-1$ $\rho(1)$ 接近 1，表明相邻位点间的正相关系数接近 1. 类似的，可以求出 n 步相关系数为 $\rho(n) = (2p-1)^n$, 具体推导见附录 1。如果想利用相邻位点的相关性信息对某一个位点做估计，则对于给定的相关系数阀值 $\alpha$，令 $\rho(n) > \alpha$, 解出最多可以利用的相邻位点数目为 $2n = \frac{2\log(alpha)}{\log(2p-1)}$

## 2 仿真结果

使用 R 语言产生一 Markov Chain, 其中 p=0.95, 由于过程平稳，可利用相关函数的遍历性质求出 R(n) 的样本值，将其与理论结果进行比较，作图如下：




## 3 模型检验

对于已知结构的 RNA 序列，试图求出 p, 使得均方误差 $\sum_{k=1}^{n}((2p-1)^k - \hat{R}(k))^2$ 最小，其中 n 取 $\lfloor \frac{\log(\alpha)}{\log(2p-1)} \rfloor$, 其中 $\alpha$ 事先给定。

实际对已知结构（链长为 300）的某条 RNA 求相关系数发现 $\hat{R}(1) = 0.46$ 且当 $n \geq 3$ 时 $\hat{R}(n) \leq 0$, 这说明之前关于相邻链的相关性的假定很可能是不正确的，需要寻找新的可以更好地推断原 RNA 二级结构的数学模型。

## 4 Appendix 1: $\rho(n)$ 表达式的推导

假设 $X_0$ 服从 Bernoulli 01 分布，概率为 $\frac{1}{2}$，记为行向量 $\vec{p_0} = (\frac{1}{2}, \frac{1}{2})$, 第一个位置表示处于状态 0 的概率，第二个位置表示处于状态 1 的概率。则 $X_1$ 的分布为 $\vec{p_0} = \vec{p_1}P = \vec{p_0}$, 递推得到 $X_n$ 的分布为与 $\vec{p_0}$ 相同。在这种情形下，$R(1) = P(X_n = 1, X_{n-1} = 1) = P(X_{n-1} = 1)P(X_n =$

$1|X_{n-1}=1)=\frac{p}{2}$。为计算 n 步自相关函数，需要先求出 n 步转移矩阵 $P^n$ 的表达式，为此可以采用特征值分解的方法，先将矩阵 P 分解为:

$$P=Q\begin{pmatrix}1 & 0\\ 0 & 2p-1\end{pmatrix}Q^{-1}, Q=\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}$$

则

$$P^n=Q\begin{pmatrix}1 & 0\\ 0 & (2p-1)^n\end{pmatrix}Q^{-1}=\begin{pmatrix}\frac{1+(2p-1)^n}{2} & \frac{1-(2p-1)^n}{2}\\ \frac{1-(2p-1)^n}{2} & \frac{1+(2p-1)^n}{2}\end{pmatrix}$$

于是 $R(n)=\frac{1+(2p-1)^n}{4}, \rho(n)=\frac{R(n)-\frac{1}{4}}{\frac{1}{4}}=(2p-1)^n$

## 5 参考文献

# References

[1] 生物信息学