Modeling Molecular Evolution

Ideas to Review

Probability: Simple, Joint, Conditional Probabilities

$P(H)$, $P(\text{sum of 2 dice} = 3)$, $P(\text{sum of 2 dice is 3} \mid \text{first roll is 1})$

$P(A \text{ and } B)$ versus $P(A \mid B) = \dfrac{P(A,B)}{P(B)}$

Exponentials:

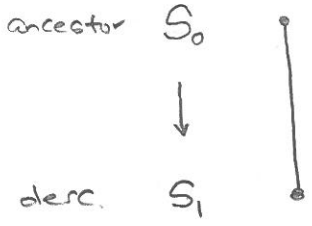
$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$ Taylor Expansion

$p'(t) = cp(t) \qquad p(0) = 5$ "Exponential Growth" differential equation

Matrix Multiplication:

Simulation of sequences on trees ...

I) Modelling Evolution on a 1-edge tree:



Consider a 2-state model

R = purine A, G

Y = pyrimidines C, T

We need a 1) ROOT DISTRIBUTION $p_0 = (p_R, p_Y)$

with $p_R = P(\text{PURINE at } S_0) = P(S_0 = R)$

$p_Y = P(S_0 = Y)$

This is a probabilistic distribution so $p_Y, p_R \geq 0$, $p_Y + p_R = 1$.

Example: $p_0 = (.7, .3)$ more likely to see R than Y in S_0

$= (.5, .5)$ R, Y equally likely

We also need 2) a MARKOV TRANSITION M $= M_{01}$

with entries transition probabilities of various state changes from S_0 to S_1

$M =$ columns S_1: descendant

		$S_1 = R$	$S_1 = Y$
rows S_0: ancestor	$S_0 = R$	$P(S_1 = R \mid S_0 = R)$	$P(S_1 = Y \mid S_0 = R)$
	$S_0 = Y$	$P(S_1 = R \mid S_0 = Y)$	$P(S_1 = Y \mid S_0 = Y)$

We often simplify notation $p_{YR} = P(S_1 = R \mid S_0 = Y)$

(S_0 ↑, S_1 ↑)

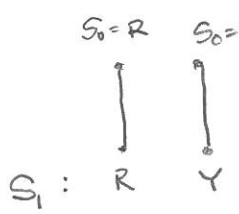
$$M = \begin{pmatrix} p_{RR} & p_{RY} \\ p_{YR} & p_{YR} \end{pmatrix}$$

Note:

1) the entries of M are conditional probabilities (transition probs.)

2) the rows of M sum to 1

Row 1: Given that $S_0 = R$, $P(S_1 = R \mid S_0 = R) + P(S_1 = Y \mid S_0 = R) = 1$



States that S_1 is either in state R with prob P_{RR}

or " " " " Y " " P_{RY}

Row 2 similarly.

Taken together, $(\vec{p}_0, M)$ are parameters of our substitution model.

Example: $\vec{p}_0 = (.5, .5)$ $\quad M = \begin{pmatrix} .8 & .2 \\ .1 & .9 \end{pmatrix}$ $\quad P_{RR} = .8 \quad P_{RY} = .2 \quad P_{YR} = .1 \quad P_{YY} = .9$

S_0:

DEMO

S_1:

Using $(\vec{p}_0, M)$, we can generate aligned sequences of length 1

or by assuming an

INDEPENDENT ↑ each site pattern generated independently

and

IDENTICAL PROCESS ↑ use same parameters $(\vec{p}_0, M)$ for each site

$\equiv$ i.i.d. = independent and identically distributed

we can generate aligned sequences of length n by repeating this process n times

Review: $(\vec{p}_0, M)$ $\Big|_{S_1:}^{S_0:}$

iid

etc

Problem: (Small Eg.) Suppose $n = 10$ sites have been generated by an iid Markov model on $\downarrow_{S_1}^{S_0}$. Give the best estimates for $(\vec{p}_0, M)$ that you can

S_0: RRYRRYYRRR

S_1: RR RY RY YRYY

Answer: $\hat{\vec{p}}_0 = (.7, .3)$

$\hat{M} = \begin{pmatrix} 4/7 & 3/7 \\ 1/3 & 2/3 \end{pmatrix}$

$\hat{}$ = estimate from data.

Part 2: Given the JOINT FREQUENCY ARRAY

$$F = \begin{pmatrix} P(S_0 = R, S_1 = R) & P(S_0 = R, S_1 = Y) \\ P(S_0 = Y, S_1 = R) & P(S_0 = Y, S_1 = Y) \end{pmatrix} = \begin{pmatrix} 4/10 & 3/10 \\ 1/10 & 2/10 \end{pmatrix}$$

$\uparrow$ P used typically ...

Note the entries of F add to 1, F is called the PATTERN FREQUENCY ARRAY, JOINT DISTRIBUTION of STATES at LEAVES of TREE, etc

Its entries are nonnegative and sum to 1.

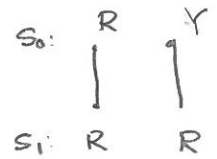
Part 3a) Given the distribution of states at S_1

$\hat{p}_1 = (.5 \quad .5)$

3b) ... Do this theoretically ...

3b) Given $\vec{p}_0 = (.7 \quad .3)$ $\quad M = \begin{pmatrix} 4/7 & 3/7 \\ 1/3 & 2/3 \end{pmatrix}$ model parameters,

$\vec{p}_1 = \left(P(S_1 = R) \;,\; P(S_1 = Y) \right)$



$$P(S_1 = R) = P(S_0 = R)\,P(S_1 = R \mid S_0 = R) + P(S_0 = Y)\,P(S_1 = R \mid S_0 = Y)$$

$$= \left(\tfrac{7}{10}\right)\left(\tfrac{4}{7}\right) + \left(\tfrac{3}{10}\right)\left(\tfrac{1}{3}\right) = .4 + .1 = .5 \;!$$

Even better: $P(S_1 = R) = (.7 \quad .3)\begin{pmatrix} 4/7 \\ 3/7 \end{pmatrix} = (\vec{p}_1 -)\begin{pmatrix} \uparrow \\ \text{Col 1} \\ \downarrow \end{pmatrix}$

$$P(S_1 = Y) = (.7 \quad .3)\begin{pmatrix} 3/7 \\ 2/3 \end{pmatrix} = .5$$

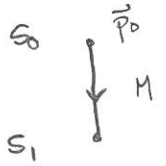
$$= P(S_0 = R)\,P(S_1 = Y \mid S_0 = R) + P(S_0 = Y)\,P(S_1 = Y \mid S_0 = Y)$$

$$= P_R P_{RY} + P_Y P_{YY}$$

Putting this together

$$\vec{p}_1 = P(S_1 = R, S_1 = Y) = \vec{p}_0 M$$ vector-matrix product

Summary 1-edge model:



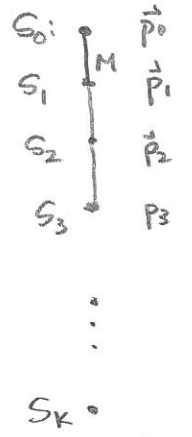
To get the distribution of status at S_1, compute $\vec{p}_1 = \vec{p}_0 M$

Current assumptions: To simulate sequences of length $n > 1$, assume an i.i.d process.

Discrete Process: Model from endpoint S_0 to endpoint S_1

$\equiv$ 1 time step

Question: How can you modify this for $K=2,3,\ldots$ time steps?



$\Delta t = 1$ time step

$\vec{P}_1 = \vec{P}_0 M$

$\vec{P}_2 = \vec{P}_1 M = \vec{P}_0 M^2$

$\vec{P}_3 = \vec{P}_0 M^3$

$\vec{P}_K = \vec{P}_0 M^K$