7.36/7.91 recitation

2-19-2014
CB Lecture #4
Announcements / Reminders

Homework:
- PS#1 due Feb. 20th at **noon**.
- Late policy: ½ credit if received within 24 hrs of due date, otherwise no credit
- Answer key will be posted 24 hrs after due date

Project:
- Teams, Title, 1 paragraph summary due Tuesday Feb. 25
- Teams of 1-5 people unless approved by instructor
Basic Linear Algebra Review

- way to compactly represent and operate on sets of linear equations:

\[ 2x_1 + 4x_2 = 10 \]
\[ -5x_1 + x_2 = -3 \]

where

\[ \vec{x} = (x_1, x_2) \]
\[ A = \begin{bmatrix} 2 & -5 \\ 4 & 1 \end{bmatrix} \]
\[ \vec{b} = (10, -3) \]
Basic Linear Algebra Review

Simple operations:
- Dot product of two row vectors \( \vec{x} = (x_1, x_2, x_3) \quad \vec{y} = (y_1, y_2, y_3) \)
  \[
  \vec{x} \cdot \vec{y} = (x_1, x_2, x_3) \cdot (y_1, y_2, y_3) = x_1y_1 + x_2y_2 + x_3y_3
  \]
- Matrix multiplication:
  \[
  A = \begin{bmatrix}
  \vec{a}_1 \\
  \vec{a}_2 \\
  \vdots \\
  \vec{a}_m 
  \end{bmatrix} \quad B = \begin{bmatrix}
  \vec{b}_1^T \\
  \vec{b}_2^T \\
  \vdots \\
  \vec{b}_p^T 
  \end{bmatrix} \quad \Rightarrow \quad A \times B = \begin{bmatrix}
  \vec{a}_1 \cdot \vec{b}_1 & \vec{a}_1 \cdot \vec{b}_2 & \cdots & \vec{a}_1 \cdot \vec{b}_p \\
  \vec{a}_2 \cdot \vec{b}_1 & \vec{a}_2 \cdot \vec{b}_2 & \cdots & \vec{a}_2 \cdot \vec{b}_p \\
  \vdots & \vdots & \ddots & \vdots \\
  \vec{a}_m \cdot \vec{b}_1 & \vec{a}_m \cdot \vec{b}_2 & \cdots & \vec{a}_m \cdot \vec{b}_p
  \end{bmatrix}
  \]
  - Note that inner dimensions must agree:
    If \( A \in \mathbb{R}^{m \times n} \) and \( B \in \mathbb{R}^{n \times p} \) then \( A \times B \in \mathbb{R}^{m \times p} \)
Markov Models (Chains)

- Defined by a set of $n$ possible states $s_1, \ldots, s_n$ at each timepoint.

- **Markov property:** Transition from state $i$ to $j$ (with probability $P_{i,j}$) depends *only* on the previous state, not any states before that. In other words, the future is conditionally independent of the past given the present:

$$P(S_{t+1} = k|S_1 = s_1, \ldots, S_t = s_t) = P(S_{t+1} = k|S_t = s_t)$$

**Example:** if we know individual 3’s genotype, there’s no additional information that individuals 1 and 2 can give us about 5’s genotype

- Probability of having a class having an exam that week

  - *not Markov:* prob. of having an exam during a week influenced by events further back than just 1 week (if there was an exam 2 weeks ago, likely not an exam this week)

- Board games whose moves are entirely determined by dice

  - **Markov:** prob. of future event depends only on the current board and outcome of dice roll
Markov Chains

- Instead of realizing a set of states (one particular state with probability 1 and all others with probability 0 at each timepoint), we can model more general processes by defining a probability distribution over states at each timepoint:

\[ \vec{q}^t = (q_1, \ldots, q_n) \quad 0 \leq q_i \leq 1, \sum_{i=1}^{n} q_i = 1 \]

- Probability distribution changes over time according to transition matrix \( P \)

\[ \vec{q}^{t+1} = \vec{q}^t P \]
\[ \vec{q}^{t+k} = \vec{q}^t P^k \]
Markov Models (Chains)

- Defined by a set of $n$ possible states $s_1, ..., s_n$ at each timepoint.

- **Markov property:** Transition from state $i$ to $j$ (with probability $P_{i,j}$) depends *only* on the previous state, not any states before that. In other words, the future is conditionally independent of the past given the present:

$$P(S_{t+1} = k | S_1 = s_1, ..., S_t = s_t) = P(S_{t+1} = k | S_t = s_t)$$

- Size and constraints on transition matrix $P$:
  - Size: $n \times n$
  - Interpretation: From current state $i$, you must end up in some state $j$ after transition

$$P = \begin{bmatrix}
P_{1,1} & P_{1,2} & \cdots & P_{1,n} \\
P_{2,1} & P_{2,2} & \cdots & P_{2,n} \\
\vdots & \vdots & \ddots & \vdots \\
P_{n,1} & P_{n,2} & \cdots & P_{n,n}
\end{bmatrix}$$

$$\sum_{j=1}^{n} P_{i,j} = 1 \quad \forall i$$
Markov Models (Chains)

- Defined by a set of n possible states $s_1, ..., s_n$ at each timepoint.

- **Markov property:** Transition from state $i$ to $j$ (with probability $P_{i,j}$) depends *only* on the previous state, not any states before that. In other words, the future is conditionally independent of the past given the present:

$$P(S_{t+1} = k | S_1 = s_1, ..., S_t = s_t) = P(S_{t+1} = k | S_t = s_t)$$

- If at time $t$ the probability distribution over the $n$ states is

$$\vec{q}^t = (q_1^t, q_2^t, ..., q_n^t)$$

what is the probability of being in state $i$ at time $t+1$?

$$q_{i}^{t+1} = \sum_{s=1}^{n} q_{s}^{t} P_{s,i}$$
Markov Chains

- If all entries of $P$ are strictly positive ($P_{i,j} > 0$), there is a “stationary” (or “limiting”) distribution in the limit of infinite time:

$$\lim_{t \to \infty} \vec{q}^t = \lim_{t \to \infty} \vec{q}P^t = \vec{r}$$

- The stationary distribution satisfies: $\vec{r} = \vec{r}P$

- Since all entries of distribution must sum to 1, can set up system of eqns to solve:

$$\vec{r} = (r_1, r_2, \ldots, 1 - \sum_{i=1}^{n-1} r_i)$$

- May also notice that $\vec{r}$ is an eigenvector of $P$ with eigenvalue 1. Can use eigenvector approaches instead of systems of eqns to determine $\vec{r}$ if you’re familiar with those
Practice Problem

- You decided to make a model of purine (R) and pyrimidine (Y) evolution. Multiple sequence alignment of promoters (50% R, 50% Y) leads to:

\[
PAM_1 = \begin{bmatrix} 0.995 & 0.005 \\ 0.015 & 0.985 \end{bmatrix}, P_{R,R} = 0.995, P_{R,Y} = 0.005, P_{Y,R} = 0.015, P_{Y,Y} = 0.985
\]

- What is the composition of a sequence evolving under this model after a long time?

Let \( P_Y = 1 - P_R \):

\[
(P_R, 1 - P_R) = (P_R, 1 - P_R) \begin{bmatrix} 0.995 & 0.005 \\ 0.015 & 0.985 \end{bmatrix}
\Rightarrow (P_R, P_Y) = (0.75, 0.25)
\]

- What is PAM\(\infty\)?

Since the (0.75, 0.25) outcome must be the same no matter where we start from (e.g. (1, 0) or (0, 1)):

\[
\lim_{t \to \infty} PAM_t = \begin{bmatrix} 0.75 & 0.25 \\ 0.75 & 0.25 \end{bmatrix}
\]
Practice Problem

- You decided to make a model of purine (R) and pyrimidine (Y) evolution. Multiple sequence alignment of promoters (50% R, 50% Y) leads to:

\[
PAM_1 = \begin{bmatrix} 0.995 & 0.005 \\ 0.015 & 0.985 \end{bmatrix}, P_{R,R} = 0.995, P_{R,Y} = 0.005, P_{Y,R} = 0.015, P_{Y,Y} = 0.985
\]

- What would be the average % sequence identity between an initial sequence (composition 50% R, 50% Y) and the sequence evolved from this initial sequence under the PAM\(\infty\) matrix?

- In PAM\(\infty\), \(P_{R,R} = 0.75\) and \(P_{Y,Y} = 0.25\). We start with 50% R and 50% Y. The fraction of R’s remaining the same is \((0.75)(0.50) = 0.375\) and the fraction of Y’s remaining the same is \((0.25)(0.50) = 0.125\). Therefore the total % sequence identity is \(0.375 + 0.125 = 0.50\) or 50%. 
- Evolutionary time measured in Percent Accepted Mutations (PAMs)

- One PAM: 1% of the residues have changed, averaged over all 20 amino acids.

- To get the relative frequency of each type of mutation, count the times it was observed in a database of multiple sequence *global* alignments

- The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence

- Mutation frequencies assume a Markov model of evolution. Other matrices derived from PAM1: PAM250 ~ \((\text{PAM1})^{250}\)

- BLOSUM matrices are based on *local* alignments

- BLOSUM 62 is a matrix calculated from alignment of sequences with ~62% identity.

- BLOSUM matrices are based on observed alignments; unlike PAM, they are not extrapolated from comparisons of closely related proteins

- BLOSUM 62 is the default matrix in BLAST. It’s tailored for comparisons of moderately distant proteins.

- Alignment of more distant proteins may be more accurate with a different matrix based on substitutions observed in more distantly evolved proteins

Jukes-Cantor model

- the number of observed differences between two homologous sequences is smaller than the actual number of changes that have occurred, due to reversions (e.g. $A \rightarrow G \rightarrow A$)
- can underestimate the genetic distance between the sequences
- How to compensate? need some model of how mutations occur

- Jukes-Cantor model assumes that all mutations are equally likely and occur with rate $\alpha$; if this is true, then you can apply the following correction:

$$K = -\frac{3}{4} \ln \left[ 1 - \frac{4}{3} P \right]$$

$P = \text{observed fraction sites that differ}$
$K = \text{actual number of substitutions}$

- This is very simple; other models are much more complex (e.g. Kimura, which has transitions $C \leftrightarrow T$ and $A \leftrightarrow G$ occurring more frequently than transversions $R \leftrightarrow Y$ and $Y \leftrightarrow R$).
Positive / Negative Selection

$K_a/K_s$ (or dN/dS) test:
- $K_a$ (or dN): # of nonsynonymous substitutions per nonsynonymous site
- $K_s$ (or dS): # of synonymous substitutions per synonymous site

- What are typical $K_a/K_s$ ratios you expect for protein-coding genes?
  - Most proteins have evolved to a near-optimal sequence & structure, so most mutations will be deleterious ($K_a/K_s << 1$).

  The frequency of different values of $K_a/K_s$ for 835 mouse–rat orthologous genes.


$K_a/K_s \approx 1$ generally means neutral evolution (averaged over calculated region) - e.g. pseudogenes

$K_a/K_s > 1$ generally means positive selection (e.g. immune system genes coevolving with parasites) - see first 3 pages of Sabeti review on Positive Selection in humans in “Resources” on course website