Sarah and Isabella,

Your paper and presentation on the Jukes-Cantor model and relations with algebraic statistics were good although there were some (relatively small) inaccuracies and also some things I would have liked to see explained in more detail to show you really understood what was going on. Two of the difficult things about projects like this are

1) Learning new things as necessary (and I guess you needed to do a lot of that here).
2) Reconciling the information you get from different sources and understanding the terminology and notation they use.

I think there were several times when similar notation used for different things in different sources confused you. For instance on page 6, the $\pi_A, \pi_C, \pi_G, \pi_T$ should not be transition probabilities. The quantities that are all equal are (by assumption), the distribution of the nucleotides over the whole genome. In other words, if you pick a random location in the genome, the probabilities of seeing A, C, G, T at that location are all the same.

It would have been good to reorganize what you say and include some more details about the exact relation between the JC rate matrix from page 7 is related to the transition matrix from page 6. They are equivalent in a sense, but I think it's probably better to start from the rate matrix Q. The idea is that $\theta(t)$ has to be a solution of the differential equation $\theta'(t) = Q \cdot \theta(t)$ with $\theta(0) = I$ (identity matrix). Then, as you said, $\theta(t) = e^{Qt}$, and $\theta(t)$ is given as in the equation on your page 7. The point though, is that the entries in $\theta(t)$ are the transition probabilities per unit time μ_i and π_i from before (you need the i's if you're looking at more than one site in the genome, but otherwise, they could be omitted as below). Hence

$$\mu = \frac{1}{4}\left(1 + 3e^{-4\alpha}\right)$$

and

$$\pi = \frac{1}{4}\left(1 - e^{-4\alpha}\right)$$

from setting $t = 1$ in the formula for $\theta(t)$. (Note that there's a small error in the formulas at the top of page 8.)

The expected number of mutations over a time period t is $3\alpha t$ as you say. But since you have set up all the equations, it would also have been good to say more about the actual derivation of the JC distance (or "correction"). The idea is that starting from any mother (ancestor) species, the probability of a change after time t is the sum of the three off-diagonal terms in the corresponding row of the $\theta(t)$ matrix: $\frac{3}{4}(1 - e^{-4\alpha t})$. So if you see that k/n of the positions in a sequence of the genome have changed, then the maximum likelihood estimate for the branch length is obtained just by solving the equation

$$\frac{k}{n} = \frac{3}{4}(1 - e^{-4\alpha t})$$

for $-4\alpha t$, then multiplying that by $\frac{-3}{4}$ to get the $3\alpha t$:

$$\frac{4k}{3n} = 1 - e^{-4\alpha t},$$

so

$$-4\alpha t = \ln\left(1 - \frac{4k}{3n}\right)$$

and then multiply both sides by $\frac{-3}{4}$.

It would have been good to say more about where the p_{123}, p_{dis}, p_{ij} formulas come from in the paper (as you did in the talk).

Also see the project paper hardcopy for some more specific comments.

Final Project Presentation: 92 (A-)

Final Project Paper: 92 (A-)