Ty,

Thanks for being a good sport and taking all of my suggestions for your oral presentation to heart. You did a good job there and I think people got a better idea of what algebraic statistics is about by the example you did of the 3×3 independence model and the MLE for the mixture model. Unfortunately, the paper just is not at that same level because there are a number of things you either got more or less wrong or did not explain very well. You also didn't include the second example from the talk, which would have made a really nice addition!

In the classes you have taken with me, I have noticed a consistent pattern in how you approach learning mathematics along the lines of what I think happened with this paper. This is something you will need to change if you want to do anything related to this subject in the future. What I mean is that you too often "settle for" superficial understanding of topics or subjects without really making yourself work out all of the details or forcing yourself to think things through carefully. See the comment 7 below for the biggest single example of this from the paper. But that's far from the only example. I have seen the same thing on problem sets, in oral problem presentations, on the midterm exam, and so forth. When I have seen students with similar issues in the past, it has often been because they didn't really care and just couldn't be bothered to put in the time to really learn the material. I DON'T think that is the case with you, though, so I have to say that I'm somewhat mystified why this keeps happening in the work you submit. (Could it be that math always came easily for you in the past and you never had to do this before?) The thing I will say is that I think you need to learn to be more self-critical and hold yourself to a higher standard. I wouldn't say this if I wasn't convinced you *had the ability and motivation* to do that.

Specific Comments

1. I guess you are thinking of thinks like the Captcha tests built in to lots of website logins. That doesn't seem like such a big inconvenience, though. And if it adds a layer of protection against bot attacks on websites that have information I want to keep private (like credit card numbers, etc.) I'm willing to put up with them.
2. This is not very precise. I think you are trying to suggest the properties of Markov chains that you introduce two paragraphs ahead. But it would be better to wait and be more careful there.
3. It would have been good to reorganize what you say and include some more details about the relation between what is often called the JC *rate matrix*:

$$Q = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

and the substitution matrix from page 4. 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). This is a more precise way of saying that Jukes-Cantor is based on the assumption that the time evolution is a Markov process. Any continuous-time memoryless dynamical system looks like this. Then, by general properties of solutions of linear constant-coefficient ODE, $\theta(t)$ is given by the matrix exponential:

$$\theta(t) = e^{Qt},$$

and $\theta(t)$ is the matrix in the equation on your page 4. 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 $\theta(t)$.

4. The subscripts in the μ_i and π_i allow the transition probabilities to be different in different locations along the DNA sequence. In the basic Jukes-Cantor model, they are the same for all transitions between nucleotides at any particular site as shown in the diagrams from page 6. You only need the subscript i's if you are talking about different sites in the genome. Different nucleotides (A,C,G,T) *are not the same thing as different sites.* To talk about the *sites* in DNA sequences, there has to a way to align those sequences and identify one pair of nucleotides in one sequence with one pair of nucleotides in another sequence.
5. I wouldn't say the Poisson distributions are "uniquely reliable." They are a standard way to model the number of occurrences of some event within limited regions in continuous time and/or space.
6. There is a typo here – the condition in the first term should be $rt = C$, not $rt = A$.
7. I'm afraid you really misunderstood the example from David Cox's survey that you are working from here. The thing you didn't notice or didn't understand is that this is supposed to be a *"binary"* Jukes-Cantor model. That means that π_0, π_1 are just binary data: 0 or 1. *In other words, this whole discussion does not apply directly to non-binary data like the A, T, G, C nucleotides in DNA.* That's why the transition matrices on the three branches are just 2×2 matrices and not 4×4 matrices. Also, this set-up is different from the Jukes-Cantor DNA model because notice that the three transition matrices *could be different.* The discussion of the variety corresponding to this statistical model, following Cox, is good. The problem is that it just doesn't apply to the situation you are studying in the rest of the paper. The thing that is parallel in the DNA situation is the derivation of the "condensed" Jukes-Cantor model from pages 150 and 151 of the Pachter-Sturmfels book. I don't think you thought about this carefully enough to really understand what was going on and recognize that what you said could not apply to the DNA evolution model.

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 $\frac{-3}{4}$.

Final Project Presentation: 92 (A-)

Final Project Paper: 82 (B-)