Notes on Tseng & Liang 2005 Mol. Bio. and Evo.

I have a matrix and from that I got **.**

This is a matrix of transition probabilities after a time *t*.

However, BBTM is a matrix of similarity scores, . These similarity scores are time-dependent, which is why my BBTM alignments were so bad.

I don't know what λ is. I don't think it matters, though, I once multiplied a substitution matrix by 10,000 and got the same result. I'll set it by trying to reproduce BBTM. I can set it by matching one entry, so I have the remaining 399 entries to check that I set it correctly. It also means the base of the logarithm doesn't matter.

In Jimenez-Morales & Liang 2011 PLoS One, π refers to a probability distribution, and it's basically used like a probability operator.

In Tseng and Liang 2005 Molecular Biology and Evolution, it is said:

For node and node separated by divergence time , the time reversible probability of observing residue in a position at node and residue of the same position at node is:

The way I'd translate this into the notation I'm used to is to consider a continuous markov process , and let and denote times, and :

is the same for all , so it's actually only a function of the state: I think that's what means. Since is time-homogenous, depends only upon , and I think it's represented by .

My interpretations of the symbols seem to make sense. So back to the equation for the elements of .

What's in the logarithm looks like a conditional probability over the prior probability of what's left of the conditional bar: a puf!

So, this is like, a logarithmic "strength of evidence" thing, for the degree of confirmation that the original residue was residue , at time 0, after observing that it is at time . OR SO IT SEEMS. in the Tseng/Liang paper they do a substitution:

where " is the joint probability of observing both residue time and at two nodes separated by time ".

So, on the bottom of that fraction, there's , the probability that those two sequences would have those two residues at that position if they were actually entirely unrelated (assuming the only kind of evidence you admit is evidence from evolutionary relation: for example, if you know it's an active site in both proteins, then what you find at one is evidence about what you will find at the other, sinceit tells you something about what active sites are like).

On the top of the fraction, there's , the probability that these two sequences would have those two residues at that position if they actually *are* related.

So the numerator and the denominator are the probabilities of the same conjunction given two different models: the ratio is a Bayes factor.

I already knew that . But I never thought of the puf before as a Bayes factor between X and a model in which A and B are independent, but which assigns the same prior probabilities as X. BUT IT IS! Wow. That's going in the Bayes book tonight.

So, if you've got , then these members of are what you add to it to update. And it kind of makes sense that that's what a sequence alignment program would do? Try align positions where it seems likely that they are related at a distance , or at least that that is many times more likely than the null hypothesis?

SO. What I'm missing, to derive , is the function of amino acid types . Which, I guess, is just the amino acid frequencies in their dataset. WHERE ARE THEY.