It seems to me that the p matrices David sent me are not derived from the published q matrix.

However, I have a problem with both the q matrix and the p matrix that I'm using. They don't seem to really *be* for a time unit of 1.

One evolutionary time unit means the time in which you have an expected 1 change per 100 residues.

But neither my matrices nor his seem to have this property.

Here's me calculating expected number of changes per hundred residues for my P(1) and his:

his\_p = parse\_david('MTMout1.p')

my\_p = transition\_probability\_matrix(qout, 1)

**print**('Average rate of change of a hundred residues ' \

+' per evolutionary time unit, given the '\

+ 'rates in qout:')

**print**(avg\_rate(qout, pi\_out) \* 100)

**print**('')

**print**('Expected number of residues out of a hundred different '\

+ 'after one evolutionary time '\

+ 'unit, given David Jimenez-Morales\'s version of P(1):')

**print**(expected\_changes(his\_p, pi\_out) \* 100)

**print**('')

**print**('Expected number of residues out of a hundred '

+ 'different after one evolutionary time '\

+ 'unit, given my version of P(1):')

**print**(expected\_changes(my\_p, pi\_out) \* 100)

And the output:

Average rate of change of a hundred residues per evolutionary time unit, given the rates in qout:

1.98001126631

Expected number of residues out of a hundred different after one evolutionary time unit, given David Jimenez-Morales's version of P(1):

1.743119403

Expected number of residues out of a hundred different after one evolutionary time unit, given my version of P(1):

1.93189438621

Q seems like, two times higher than it should be. So, I used .5 \*Q and got results more like I expected.

his\_p = parse\_david('MTMout1.p')

my\_p = transition\_probability\_matrix(.5 \* qout, 1)

**print**('Average rate of change of a hundred residues ' \

+' per evolutionary time unit, given the '\

+ 'rates in .5 \* qout:')

**print**(avg\_rate(.5 \* qout, pi\_out) \* 100)

**print**('')

**print**('Expected number of residues out of a hundred '

+ 'different after one evolutionary time '\

+ 'unit, given my version of P(1) relative to .5\*Q, '\

+ 'P(.5) relative to original Q:')

**print**(expected\_changes(my\_p, pi\_out) \* 100)

Output:

Average rate of change of a hundred residues per evolutionary time unit, given the rates in .5 \* qout:

0.990005633155

Expected number of residues out of a hundred different after one evolutionary time unit, given my version of P(1) relative to .5\*Q, P(.5) relative to original Q:

0.977849393937

Really I would have expected exactly 1, from both my P(1) and David-Jimenez-Morales's. If not, then what's the "evolutionary time unit" mean? When I'm figuring out which sequence identities to associate with each matrix, I can't just take it from the ones used for PAM, can I, if "evolutionary time unit" means a different amount of change?

Ooh, that's a good way to test my assumptions. Download PAM1 and see if it works the way I think it does. Or, just check if the PAM matrices used by Clustal really seem, according to my assumptions, to fit into the given seq id range.

Okay, here it is, me running the same "expected\_changes" function on a PAM1 transition probability matrix.

# Amino acid frequencies from http://www.tiem.utk.edu/~gross/bioed/webmodules/aminoacid.htm retrieved August 27 2012

pi\_ver = dict({'A': 7.4e-2,

'R': 4.2e-2,

'N': 4.4e-2,

'D': 5.9e-2,

'C': 3.3e-2,

'E': 5.8e-2,

'Q': 3.7e-2,

'G': 7.4e-2,

'H': 2.9e-2,

'I': 3.8e-2,

'L': 7.6e-2,

'K': 7.2e-2,

'M': 1.8e-2,

'F': 4.0e-2,

'P': 5.0e-2,

'S': 8.1e-2,

'T': 6.2e-2,

'W': 1.3e-2,

'Y': 3.3e-2,

'V': 6.8e-2})

pam1 = 10\*\*-4 \* parse('pam1.txt')

**print**('Expected number of changes per hundred residues after 1 PAM, '\

+ 'according to this PAM1 matrix I found by googling:')

**print**(100 \* expected\_changes(pam1, pi\_ver))

And the output:

Expected number of changes per hundred residues after 1 PAM, according to this PAM1 matrix I found by googling:

1.02483

And there you go. I am doing this right.I mean it's off in the third significant digit, but look, the amino acid frequencies I was using only had two significant digits, this is not that precise a calculation.

I think it's time to communicate my results to David. He's going to be concerned, I think. I'm concerned. I really want him to just point out some silly mistake I'm making bu tI just can't see what it could posssibly be. Maybe he gave me the wrong amino acid frequencies?

This isn't a big deal, though. It just means I can't use the same PAM time vs %identity mappings. But that's okay, because I know how to *calculate* time vs %identity mappings, I don't need to steal them from that Clustal paper.