Definitions.

*Fi* : number of occurances of amino acid *i*

*A­ij* : Number of occurances of the mutation pair of amino acids *i* and *j*

Relative frequency:

|  |  |
| --- | --- |
|  | (d1) |

Mutability:

|  |  |
| --- | --- |
|  | (d2) |

Mutation probability:

|  |  |
| --- | --- |
|  | (d3) |

Okay. M\_ij is actually an interesting and roundabout way of writing the Markovian transition probability, *from* j *to* i.

|  |  |
| --- | --- |
|  | (1) |
|  |  |
|  | (2) |

The above equation: "assuming that it does mutate from being *j*, out of all the things it could become, what's the probability that it ends up as *i*?"

(3)

Substituting (1) and (2) into (3):

|  |  |
| --- | --- |
|  | (3) |

That's not *quite* d3; it's got the subscripts on the A's switched. However, in Dayhoff/Scwartz/Orcutt:

*"We have assumed that the likelihood of amino acid X replacing Y is the same as that of Y replacing X..."*

In the notation of the probability calculus, this means that

|  |  |
| --- | --- |
|  | (4) |

And in Dayhoff et. al's notation:

|  |  |
| --- | --- |
|  | (5) |

So using a specification of (5) to sub into (3),

|  |  |
| --- | --- |
|  | (6) |

And there ya go. p\_ji should be read "probability that j mutates into i". M\_ij should be read "probabilty that i replaces j".

The consequence of this rearrangement is that *M* is column-stochastic, and you right-multiply to update a probability distribution.

Column stochasticity:

|  |  |
| --- | --- |
|  | (7) |

So if u is the uniform distribution, left multiplication gets you:

|  |  |
| --- | --- |
|  | (8) |

So when I'm trying to find steady states, I need to make sure that I'm calculating *right* eigenvectors, since if I calculate *left* eigenvectors, I'll just get the uniform distribution.

So, updates are on the right. Dayhoff says of the assumption that mutation frequencies are the same in either direction (why's it an assumption, can't she check this by looking at her trees?) that "as a consequence of this assumption, no change in amino acid frequencies over evolutionary distance will be detected". You're just as likely to observe a transition from cysteine to alanine as you are a transition from alanine to cysteine; the alanines come as often as they go in their exchanges with cysteine, and with everything else, so the frequency stays constant.

Does this mean that *all* distributions are stationary distributions? No. There's only one frequency distribution for which we know A\_ij = A\_ji holds. A\_ij is . This depends upon the initial distribution, for each i. For example, if , then A\_ij as I've defined it is zero, or A\_ji if you define it the other way, I haven't been careful about that since they're interchangable. But if , then A\_ij may also not be equal to zero. So, all this really means is that, as a consequence of the assumption, the *initial* distribution is a stationary distribution.

If the chain is totally connected, then there's only *one* stationary distribution, and I can recover the initial distribution just by finding the stationary distribution.

So, if there's only one solution to

|  |  |
| --- | --- |
|  | (9) |

Then that solution is also the initial distribution.

There are actually quite a lot of eigenvectors, corresponding to multiple eigenvalues, though.

>>> np.linalg.eigvals(M)

array([ 0.97793781, 0.98067686, 1.0000356 , 0.98255645, 0.99764694,

0.98346459, 0.99729766, 0.9970887 , 0.98453469, 0.99571345,

0.98654696, 0.98696431, 0.99463786, 0.98894918, 0.98955988,

0.9928631 , 0.99274297, 0.99120899, 0.99225484, 0.99191917])

Only one of them is 1 to four significant digits (the number of significant digits given in the matrix, though I'm not sure if it's the number of significant digits in the eigenvalues). The reason I find this confusing is because I don't see how a stochastic matrix can be multiplied by a vector and give a vector whose elements add up to something different. You can't multiply a stochastic matrix by a probability distribution and get something other thana probability distribution, can you?

My guess is that all the eigenvectors except for 1 have a mixture of positive and negative values.