CS 640 Assignment #5
Hidden Markov Models

Generalized Hidden Markov Models are defined as follows:

1. N states: $S_1,\ldots ,S_N$
2. M symbols in alphabet
3. Initial probability distribution vector of length N: $\pi = \{\pi_1 \ldots, \pi_N\}$
4. Transition probability matrix of size N x N
   where τ_{ij} is probability of transition from state i to state j
5. Emission probability matrix of size N x M
   where $e_i(c)$ probability that state i emits character c

We refer to the transition probabilities, the emission probabilities and the initial distribution vector, collectively as the parameters, designated $\lambda = (\tau_{ij}, e_i(c), \pi)$.

Let Q be the sequence of visited states: $Q = (q_1, q_2, \ldots, q_F)$
Let O be the sequence of emitted symbols: $O = (O_1, O_2, \ldots, O_T)$ (the observed sequence).

Write a generalized Hidden Markov Model that employs the forward algorithm (which is a dynamic programming algorithm) for scoring. You may hard-code in a transition matrix, emissions matrix and start probabilities. ***Your program should read in a string of any length composed of the characters {a, c, t, g} and output the score of that string, given the HMM defined below.***

Code the forward algorithm for this HMM, filling in matrix cells $\alpha_t(i)$, where t corresponds to sequence index and i corresponds to state:
1. N = 3, hidden states S_1, S_2, S_3
2. M=4 symbols in alphabet {a, c, t, g}
3. Initial probability distribution vector $\pi = \{.25, .5, .25\}$

4. Transition probability matrix τ =

	S1	S2	S3
S1	.4	.5	.1
S2	.1	.4	.5
S3	.3	.3	.4

5. Emission probabilities e =

	a	c	t	g
S1	.4	.4	.1	.1
S2	.25	.25	.25	.25
S3	.1	.1	.4	.4

Initialization: $\alpha_1(i) = \pi_i e_i(O_1)$

Iteration: $\alpha_{t+1}(i) = \sum_{j=1}^{N} \alpha_t(j) * \tau_{ji} * e_i(O_{t+1})$

Sean Eddy generalized HMMS:http://www.nature.com/nbt/journal/v22/n10/full/nbt1004-1315.html