Thomas J Matthew
10/23/15
BME 205
Coding Cost statement

A program to compute coding cost was devised to measure the information gain between orders (contexts) of hidden markov experiments applied to kmer counts. The kmer counts (of length order + 1), generated from protein sequence data, comprised both training and test sets. A log probability matrix was computed using conditional probabilities of the contexts preceding the final base of each kmer. Total coding cost, or the bit-wise information stored in a sequence, was computed for across all sequences in the test set. Average cost per character was computed as total encoding cost divided by length of the sequence. Cost per sequence was total encoding cost divided by the number of sequences, symbolized as the number of stop characters '$'.

The total encoding cost per unit length (character or sequence) increases with order, **Table 1**. This may reflect the increasing amount of information needed to convey a larger context. It may also reflect an underlying trend that bases are not significantly determined by their context. Pseudocounts, baseline counts applied to all kmers equally, was applied to provide non-zero counts. Increasing pseudocounts increase average cost per sequence since now more kmers are appearing than would otherwise, potentially reinforcing a particularly low correlation between a base and its context, **Table 2**.

Markov Model Coding Cost Summary		
	Order_0	**Order_1**
Total Encoding Cost	290.198465024	318.405297456
AVG Cost per Char	4.60632484166	5.05405234058
AVG Cost per Seq	96.7328216748	106.135099152

Table 1. Encoding cost calculated with mark_f14_1.seqs (train) and tiny.seqs (test), using default pseudocount of 1 and alphabet without wildcard characters. Usage below for order 1.
python2.7 count-kmers -o1 --alphabet=ACDEFGHIKLMNPQRSTVWY < mark_f14_1.seqs | python2.7 coding-cost tiny.seqs > out_tiny_1.seqs

Pseudocount						
Pseudo	**1**	**2**	**5**	**10**	**100**	**1000**
AVG Cost per Seq	106.13509	109.27004	117.51695	129.54222	287.98452	1622.3918

Table 2. Average coding cost per unit sequence increase with increasing Pseudocounts