Thomas J Matthew
10/27/15
BME 205
Coding Cost statement

A program to compute coding cost was devised to measure the information gain between orders (contexts) of Markov chain 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 across all sequences in the test set. Average cost per character was computed as total encoding cost divided by sum of lengths of all sequences. 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 (bits/character) decreases with order, **Table 1**. This may reflect the increasing amount of information captured in modeling each 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 does not meaningfully increase average cost per character since only the kmer distribution is smoothed to normal distribution and not the median **Table 2**.

Markov Model Coding Cost per Character (bits/character)				
Train Set	**Test Set**	**Order_0**	**Order_1**	**Order_2**
mark_f14_1.seqs	mark_f14_1.seqs	4.22738635773	4.20524103718	4.17136039622
mark_f14_1.seqs	mark_f14_2.seqs	4.22571769471	4.20490937611	4.18923050108
mark_f14_2.seqs	mark_f14_1.seqs	4.22751962386	4.20630863234	4.18925810655
mark_f14_2.seqs	mark_f14_2.seqs	4.22558404695	4.20384919047	4.17138607024

Table 1. *Encoding cost calculated with default pseudocount of 1 and alphabet without wildcard characters.*

Pseudocounts affect on Cost per Character (bits/char)						
Pseudo	**1**	**2**	**5**	**10**	**100**	**1000**
Cost/Char	4.20490937	4.20491050	4.20492588	4.20497968	4.20785892	4.24940586

Table 2. *Average coding cost per unit sequence increase with increasing Pseudocounts. Trained with mark_f14_1.seqs and tested with mark_f14_2.seqs*