Bio 131 Final Project
August Staubus

Background:

the shuffled list is found by calculating the product of the frequency of each amino acid in a given position in each possible k-mer in the sequence. The profile-most-probable k-mer from the second string is then added to the list of putative motifs. The counts are again calculated for the two k-mers now in the list of putative motifs, and this process is iterated through all sequences in the shuffled list of sequences. Once the iterations are complete, an consensus motif is calculated from the list of putative motifs by finding the most frequent amino acid at each position of the k-mer. The list of motifs is then given a score equal to the sum of the hamming distance between each putative motif and the consensus motif. Then, a new random k-mer is selected from the first sequence in the shuffled list, and again the process is iterated through all sequences in the shuffled list, an consensus motif determined, and a score calculated. If the score is greater than any score observed so far, the new list of putative motifs is saved as the current “best” motifs. The process of random k-mer selection and subsequent motif generation and scoring is repeated n times. Currently, n is coded to be proportional to the length of the first sequence in the shuffled list of sequences, but can be easily changed to another integer value. After repeating n times, the list of restriction enzyme sequences is again shuffled. The shuffling process and subsequent steps are repeated i times, where i is an integer, and the motif ensemble with the highest score from all i runs is reported, along with the consensus motif and score. The algorithm is summarized in figure 1.

This algorithm also outputs a plot of the highest motif score as a function of the number of times the algorithm is run (i). This plot may be used as a diagnostic to determine whether i and

Results and Discussion:

Because the algorithm requires a length of k-mer to look for as an input, this algorithm is not particularly adept at identifying motifs of variable length. Unfortunately, the length of the PD…D/EXK motif widely variable, with the number of residues separating the PD and D/EXK regions varying from under 10 to over 40 [1]. To overcome this difficulty, I tried searching for 2-mers (in an effort to identify the PD motif) and 3-mers (to find the EXK motif) independently. However, even after setting i to 100, the algorithm returned motifs of variable score and sequence for both 2-mers and 3-mers (Figure 2) and only rarely would the motifs include the desired sequences (Figure 2). Many motifs have equivalent scores to either PD or D/EXK. It would appear that this motif is too poorly conserved to be identified using this computational method.

Indeed, further investigation revealed that the PD…D/EXK motif is more weakly conserved than I was lead to believe (Figure 3). The PD motif is not strictly conserve; often the P is absent. Additionally, the EXK is often a EXXK motif, further frustrating my efforts to identify the motif.

There is a rich history of those more qualified than I developing and refining algorithms meant to perform virtually t