**Problem Set 3**

**Problem 1**

(a) See code.

(b) See results files data<N>\_results.txt

**Problem 2**

(a) See evosig\_results.txt

(b) *Are there biases in the sequence properties of either class? If so, where does this bias come from?*

Most frequent k-mers appear to be AT rich. Most conserved k-mers appear to have a more even distribution of all four bases.

*Which of these two lists should we use to direct further our inquiry into yeast transcription factor binding sites?*

Conservation. Even though not all conservation islands will be motifs (many are due to chance or non-motif conservation), we know that individual motif instances are preferentially conserved. You would want to look at conservation in:

1. Intergenic regions
2. Intergenic vs coding regions
3. Upstream vs downstream regions

You would then want to compare conservation to controls.

*From a scan of motifs yeast\_motifs.txt, which known motifs does your scan of 6-mers find?*