Problem Set 3: Networks and Gene Regulation

Ethan Sherbondy

October 29, 2012

1 Bayesian Decision Theory

(a) The expected loss is the sum of the likelihoods all possible labelings of the data given that the actual class is k times the corresponding loss value for each (mis)classification. Symbolically, this can be expressed as:

$$\begin{aligned} E_k[L] &= \sum_j L_{kj} P(j|d) && (1) \\ &= \sum_j L_{kj} \frac{P(d|j)P(j)}{P(d)} && (2) \end{aligned}$$

(b) If the loss matrix is $L_{kj} = 1 - I_{kj}$, then: $E_k[L] = \sum_j \frac{P(d|j)P(j)}{P(d)}$, where $j \neq k$. This is the sum of the conditional probilities for each class.

(c) The loss matrix in part (b) amounts to the simple Bayes classifier: you just select for the most likely class (highest P) since that always corresponds to the minimum expected loss.

2 Gibbs Sampling For Motif Discovery

My Gibbs sampling algorithm is modeled after the algorithm described in Chapter 12.2 of *An Introduction to Bioinformatcs Algorithms*. It also makes use of relative entropy to weigh the random choice of a removal sequence (step 2 as described on page 413). I make use of pseudocounts for my profile matrix. I do not use the log-of-odds when calculating my l-mer probabilities, as Clojure handles extremely small decimal numbers fairly well.

My metric for convergence is a value referred to in the code as the *cutoff*. For simplicity, I say the algorithm has converged when cutoff rounds go by without an increase in the maximum-probability l-mer. I've arbitrarily set the cutoff to be 20 rounds in the current implementation, but it can easily be adjusted. If I had more time, it would be fun to explore using simulated annealing instead of my naive approach.

Despite using relative entropy, my solution seems to perform poorly when the AT/GC ratio is non-uniform. This suggests to me that I am not making proper use of the relative-entropy equation.

Regardless, below is a table summarizing the most frequent 10-mer motifs discovered across the four datasets provided:

Data 1	Data 2	Data 3	Data 4
ATTCGAATTC	GTCTACTACT	AAAAAAAAAA	AAAAAAAAAA
TCGAATTCCC	CTACTACTCA	TTTTTTTTTT	AACAAAAAAA
TTCGAATTCC	TCATATAACA	AAAAAAAAGA	TTTTTTTTTT
CGAATTCGAA	CTGTCTACTA	AAAAAAAACA	AAAAAAAAAT
AATTCGAATT	TCTCTTAAGA	TTGTATATAT	ATAAATAAAT

As you can see, the results for data 3 and 4 are likely not actual motifs.