

Algorithms in Bioinformatics 2, SoSe2007

Assignment sheet # 3

Tobias Kloeppe

April 30, 2007

This weeks assignment has its focus on finding/predicting motifs in a set of unaligned sequences

1 Implement a Gibbs Sampler (6 points)

Please download www-ab.informatik.uni-tuebingen.de/teaching/ss07/albi2/java/assign03.zip. Write a Java program `albi2.motif.runGibbsSampler` that takes as input a motif length and a fasta file containing a set of sequences and produces as output a predicted motif. The prediction should use Gibbs sampling, as explained in the script. The input file has to be a fasta file containing unaligned *DNA* sequences with no gaps. The output should be the l -mers that represent the found motif in the input file. Each l -mers should be printed on a line starting with the name of the original sequence. Finally describe and implement a method to decide when the Gibbs sampling has converged. Please run your program on all three examples (`test.fna` (with motif length 22), `test2.fna` (motif length 12) and `test3.fna` (motif length 8)). Does the method converges on all examples? If not please outline the reasons for the non-convergence. State at least two detailed examples on how the method could be improved.

2 Random background (2 points)

Suppose we are given t DNA sequences of length n , generated randomly, i.i.d..

First show: The probability that a given l -mer C occurs with up to d substitutions at a given position of a random sequence is:

$$p(l,d) = \sum_{i=0}^d \binom{l}{i} \left(\frac{3}{4}\right)^i \left(\frac{1}{4}\right)^{l-i}.$$

Then show: The expected number of (l,d) occurrences, that is, of length- l motifs that occur at least once in each of the t sequences with up to d substitutions, is:

$$E(l,d,t,n) \approx 4^l (1 - (1 - p(l,d))^{n-l+1})^t.$$

3 Expected values and significance (2 points)

Consider 20 random sequences, each of length 600. Using the above formula, please plot the expected number of (l,d) -occurrences as a function of l , for interesting values of l and d .

For which values of (l, d) would you consider an occurrence significant, for which not?

Assignments due: **Monday, May 07, 10h**