

## Bioinformatics I & II

Wayne Delpont & Sergei Kosakovsky Pond

### Practical 1: Pattern matching and sequence statistics

(a). Implement an L-mer filtering procedure to finding the highest scoring segment pair between two nucleotide sequences. Use seed size of  $L=11$ , find all exact matches of length 11 between two sequences and perform greedy extension in one direction, i.e. extend the hit to the right only if the score of the segment increases.

Use a score of 5 for matches and don't allow gaps/indels.

Return the maximum score of a segment

(b). Consider nucleotide sequences of length  $N$ . Vary  $N$  from  $2^4, 2^5, 2^6$  .. up to  $2^{16}$  ( $N=65536$ ). For a fixed  $N$ , generate 1000 pairs of random sequences, assuming that all nucleotides are equiprobable. For each pair compute the score of the highest scoring segment pair using the routine from part (a),  $M$ . Tabulate the distribution of  $M$  and find its mean. How does  $M$  change with the sequence size?

(c). Repeat the same experiment, but now simulate the first sequence of each pair randomly, and generate the second one by introducing random mutations at a rate of 0.05 per position into the first sequence – this is a crude approximation of the biological mutational process. Compare the mean of maximum scores  $M$  and the distribution of high scoring segment pairs with those from part (b).

### Optional

(e). Modify the scoring function to allow for  $k$  mismatches, where  $k=3$ . How does the length of highest scoring segment pair change with increasing values for  $k$ .

(f). Modify your code to extend in both directions.