

## Biosequence Algorithms, Spring 2005

### Exercise 3, February 15, 2005, at 12.15–14 in MT2

1. (Gusfield, Ex. 3.19) Show how to modify the wild-card matching method by replacing array  $C$  (which is of length  $m > n$ ) by a list or an array of length  $n$ , while keeping the same asymptotic running time.
2. Consider applying Shift-And to search exact occurrences of the pattern AATAAT on DNA sequences. Present the occurrence masks for the pattern, and simulate the search on the target AACATAATAAT.
3. Explain an extension of the Shift-And method to handle wild-cards efficiently (both in the pattern and in the text). Present the algorithms for preprocessing and searching, and analyse their complexity.
4. Present the suffix tree for the string  $S = \text{"OMALOMA"}$ , and explain how it would be used to locate occurrences of the patterns
  - (a) "ALA",
  - (b) "ALO", and
  - (c) "OMA"in string  $S$ .
5. (Gusfield, Ex. 6.1) Construct an infinite family of strings over a fixed alphabet, where the total length of the edge-labels on their suffix trees grows faster than  $\Theta(m)$  (where  $m$  is the length of the string). That is, show that linear-time suffix tree algorithms would be impossible if edge-labels were written explicitly on the edges.