Locating protein-creating genes by parsing DNA sequences with the Viterbi algorithm

Roger Cuscó, Matthew Sudmann-Day and Miquel Torrens

April 1, 2016

Basics

Text.

Challenge

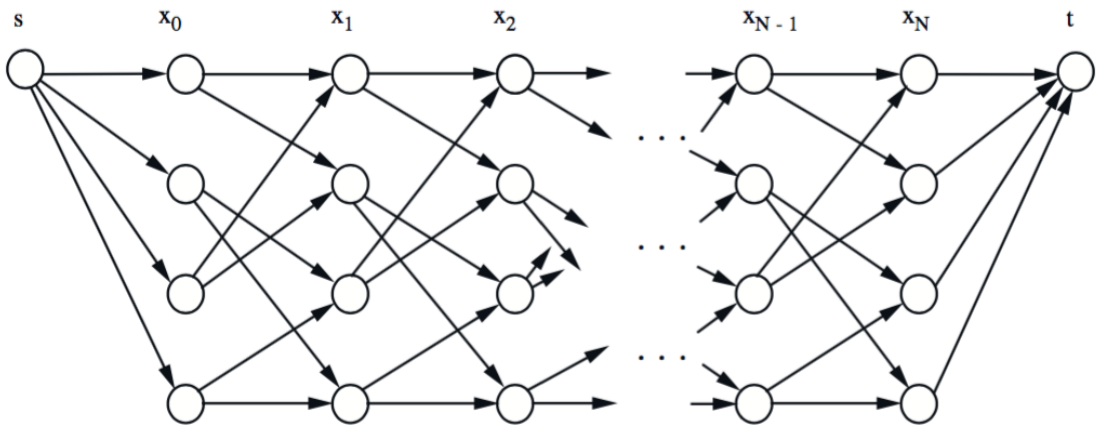
Text.

Data

- Dataset from the UCI Repository: “Molecular Biology”.
- 3190 sequences of 60 letters (“AATGCCGTAT...”).
- Each sequence is labelled as (“EI”, “IE”, “N”).
- We break down each sequence into strings of 5 letters (“AAAA”, “ACGTT”...)

Methodology

- We use a Hidden Markov Model to model the sequences.
- We predict the Hidden states using the Viterbi algorithm. Why?



Results

Text.

Conclusions

Text.