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

Roger Cuscó, Matthew Sudmann-Day and Miquel Torrens

April 1, 2016

Basics

- DNA contains nucleotides: `A`, `C`, `G`, `T`
- DNA transcription:

DNA -> pre-mRNA -> mRNA -> proteins

- Not all DNA is “expressed”

Challenge

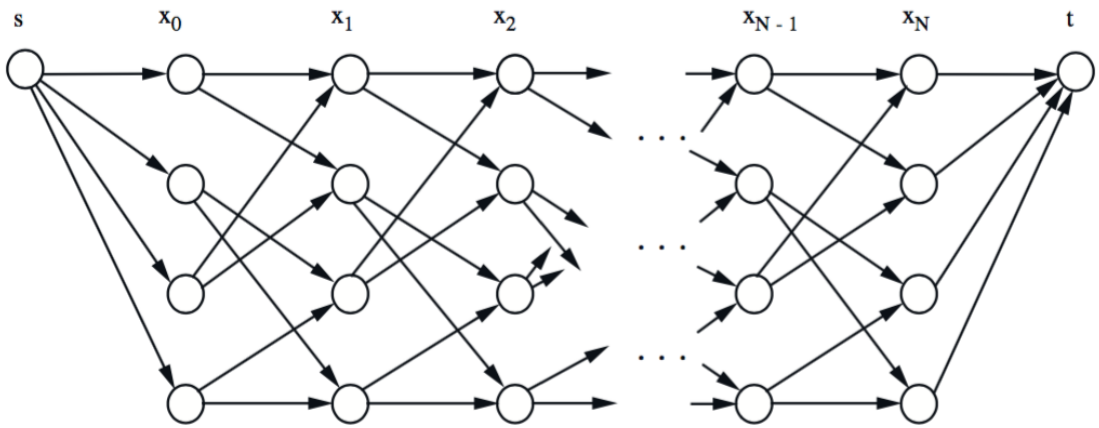
- Exons vs. Introns
- Identify transitions between exons and introns.

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

- Run Viterbi forward and backward and ensemble
- We chose 5-letter subsequence as it maximizes posterior success
- Run k-fold cross-validation but we stay with leave-one-out

Outcome:

- 82.5% in-sample success rate
- 74.1% out-of-sample success rate
  - Intron-exon transition: 69.5%
  - Exon-intron transition: 73.6%
  - Neither: 75.3%

Conclusions

1. This algorithm can reduce exponentially the regions where scientists have to look within the DNA, predicting how useful new decoded DNA sequences can be.
2. It is a key process in gene finding, disease research and drug discovery
3. Algorithm easy to train, fairly reliable, not data-hungry, computationally mild, and comprehensible to non-data-scientists
4. Current state-of-the-art techniques include neural networks and deep learning, which are more successful, but in some contexts HMM + Viterbi are still top-class!