

Graphical Programming Interface for an XML-based Hidden Markov Model Compiler

Supervisors: Gerton Lunter, Rune Lyngsoe, Jotun Hein

Hidden Markov models (HMMs) are simple yet powerful statistical models to describe sequential data. First introduced in the context of speech recognition, they are now widely used, and are particularly successful in Bioinformatics. Applications include modelling of protein domains, gene finding, probabilistic sequence alignment, and DNA binding site modelling.

Because of the size of typical biological data sets (e.g., the human genome consists of 2.85 billion base pairs), efficiency issues are crucial. On the other hand, the rate at which new insights and discoveries are published means that we'd like to rapidly test new probabilistic models, requiring a great deal of flexibility. These two requirements are often conflicting. Moreover, the algorithms themselves are known in detail, and although tedious and error-prone, implementing them is largely an 'automatic' process.

We decided to exploit this characteristic to solve the conflict between the requirements, by building a "HMM compiler", termed HMMoC. This Java program reads a high-level description of the hidden Markov model in XML format, making it very easy to change the underlying model. It produces highly efficient C++ code for the various probabilistic algorithms (Viterbi, Forward/Backward etc.). The resulting code is faster than most hand-coded implementations.

The remaining drawback of this approach is that it requires the user to learn a new (XML) language. The natural way to specify an HMM would be to draw a graph representing the Markov chain, specify transition and emission probabilities, and have the computer convert the graph into XML. This would be a valuable research tool, and would be even more valuable in teaching.

This project proposes to write such a Graphical Programming Interface, as a graphical layer between the user and the XML specification. Potentially, and if time permits, the compilation and execution-phases could be integrated, to form an integrated HMM experimentation environment. It is expected that the graphical interface will be written in Java, although other options can be discussed.

References

Early reference for HMMs in speech recognition: *A tutorial on hidden Markov models and selected applications in speech*

recognition, L.R. Rabiner, *proc. IEEE* 77, 257-286 (1989)

Introduction to HMMs for DNA sequence analysis: *Hidden Markov models in computational biology: applications to*

protein modelling, A. Krogh et al., *Journal of Molecular Biology* 235, 1501-1531 (1994)

Textbook introduction to HMMs in DNA sequence analysis, very clear and well-written: *Biological sequence analysis*,

Durbin, Eddy, Krogh and Mitchison, Cambridge Univ. Press