Logic-Statistic Models with Constraints for Biological Sequence Analysis

Christian Theil Have

Research Group PLIS: Programming, Logic and Intelligent Systems
Department of Communication, Business and Information Technologies
Roskilde University, P.O.Box 260, DK-4000 Roskilde, Denmark
cth@ruc.dk

1 Introduction and Problem Description

This project aims to investigate biologically inspired, logic-statistic models with constraints. The complexity and expressiveness of models with different kinds of constraints will be examined and algorithms to efficiently cope with inference in and training of such models will be explored. The models will be evaluated with regards to their applicability to biological sequence analysis.

Statistical models for biological sequence analysis are usually based on variants of HMMs and occasionally more expressive models like PCFGs. The size of the data often prohibits the more expressive models, so careful choice of good independence assumptions is paramount. Realistic biological models may include complicated interactions between different aspects such as codon frequencies, RNA structure and phylogenetic information. Constraints can be a way of explicitly combining aspects of such models in a more intuitive and modular way and at a higher abstraction level. The declarative nature of constraints permits a degree of freedom of implementation, allowing for potential optimizations.

Constraints are usually embedded in the model either as structure, parameters (soft, data-driven) or a combination, but can also be applied in inferencing. Consider as example a generic genefinder model, where we would like to infer the most likely sequence of *hidden* states, representing the genes and non-genes, that best explains a given *observable* sequence of nucleotides. It might be that we know that the DNA sequence for a particular organism contains at least 8000 genes, but the proposed most likely sequence has less than 8000 genes. In this case, the constraint could be used to guide the inference procedure. It has recently been suggested that this can be formulated as constraint satisfaction problem [1].

2 Background and Overview of The Existing Literature

This project is part the larger project LOgic-STatistic Modeling and Analysis of Biological Sequence Data (LoSt). Logic-statistic models can be expressed as stochastic logic programs using the PRISM language [2], which is an extension of Prolog where values of special variables are determined by random switches

rather than usual unification. PRISM includes efficient procedures for inference and parameter estimation. Stochastic logic programs can have constraints, usually in the form of equality between unified logic variables. Stochastic selection of values for such variables may lead to unification failure and resulting failed derivations must be taken into account in parameter estimation. PRISM does this using an adaptation of Cussen's FAM algorithm [3].

3 Goal of The Research

The goal is to investigate the applicability of logic-statistic models with constraints with regard to their ability to express and efficiently deal with the problems of biological sequence analysis.

4 Current Status of The Research

The research is at a very early stage where the ideas are currently being refined.

5 Preliminary Results Accomplished

A context-sensitive grammar formalism, "Stochastic Definite Clause Grammars", was implemented using PRISM and utilizing its facilities for handling failures.

6 Open Issues and Expected Achievements

I hope to find that logic-statistic models with constraints will make it easier to express complex biological models and that the achieved compositional problem structure will allow certain optimizations. I think that statistical inference and constraint solving can complement each other and that interesting techniques may be found in their intersection. Different logic-statistic frameworks may have distinct features with regard to different kinds of constraints and these features should be investigated further.

Soft constraints seems to be a nice fit to statistical models since probabilities are much like preferences or weights. In the example in section 1 we might be only 80% certain that the constraint holds. If the probability of tagging at least 8000 genes is sufficiently low, then it might be preferable to break the constraint. Using soft constraints in the context of both inference and parameter learning seems like a very interesting direction to pursue.

References

1. Petit, M., Christiansen, H.: Viterbi computation for a constrained hidden markov model. Actes JFPC (2009)
2. Sato, T., Kameya, Y.: New advances in logic-based probabilistic modeling by prism. Probabilistic Inductive Logic Programming LNCS 4911, Springer (2008) 118–155
3. Cussens, J.: Parameter estimation in stochastic logic programs. Machine Learning **44**(3) (2001) 245