Parsing Technology and RNA Folding: a Promising Start

Fabrice Lefebvre
LIX, Ecole Polytechnique
91128 Palaiseau Cedex, FRANCE
lefebvre@lix.polytechnique.fr

Abstract

The determination of the secondary structure of RNAs is a problem which has been tackled by distantly related methods ranging from comparative analysis to thermodynamic energy optimization or stochastic context-free grammars (SCFGs). Because of its very nature (properly nested pairs of bases of a single stranded sequence) the secondary structure of RNAs is well modeled by context-free grammars (CFGs). This fact has been recognized several years ago by people who used context-free grammars as a tool to discover some combinatorial properties of secondary structures. More recently SCFGs were used by several teams (esp. David Haussler's team at UC Santa Cruz) as an effective tool to fold RNAs through Cocke-Younger-Kasami-like parsers. Until 1996, and in the context of RNA folding, CFGs and their derivatives where still considered theoretical tools, barely usable outside the computer scientist lab. The exception of SCFGs seemed promising, with all the hype around Hidden Markov Models and other stochastic methods, but it remained to be confirmed for RNAs longer than 200 bases.

The main obstacle to the use of context-free grammars and parsing technology for RNA folding and other closely related problems is the following : suitable grammars are exponentially ambiguous, and sentences to parse (i.e. RNA or DNA sequences) typically have more than 200 words, and sometimes more than 4000 words. These figures are rather unusual for ordinary parsers or parser generators, because they are mostly used in the context of natural language parsing, and thus do not have to face the same computation problems. Fact is, most people dealing with RNA folding problems were manually writing dynamic programming based tools. This was the case for folding models popularized by Michael Zuker, and based on free energy minimization. This was also the case for folding models based on SCFGs. This was in effect the case for just about every computer method available to fold or align sequences. Parsing sequences was not an issue because it simply seemed too slow, too memory hungry and even unrelated.

In 1995, I showed that S-attribute grammars were perfectly able to handle both the thermodynamic model and the stochastic model of RNA folding. I then introduced a parser generator which was able, given a proper S-attribute grammar, to automatically write an efficient parser based on suitable optimizations of Earley's parsing algorithm. All generated parsers turned out to be faster and less memory hungry than other available parsers for the same exponentially ambiguous grammars and the same sequences. More surprisingly, these parsers also turned out to be faster than hand-written programs based on dynamic programming equations. This was the first proof that improvements in parsing technology may certainly be put to good use in biocomputing problems, and that they shall lead to better algorithms and tools.

While trying to overcome some limitations of SCFGs, I generalized S-attribute grammars to multi-tape S-attribute grammars (MTSAGs). The automata theory counterpart of a MTSAG would be a non-deterministic push-down automaton with several one-way reading heads, instead of a single one-way reading head as it is the case for CFG. Given these MTSAGs, a generalization of the previous single-tape parser generator was the obvious way forward.

Thanks to this new parser generator, I was able to show that most biocomputing models previously based on dynamic programming equations were unified by MTSAGs, and that they were better handled by automatically generated parsers than by handwritten programs. It did not matter whether these models were trying to align sequences, fold RNAs, align folded RNAs, align folded and unfolded RNAs, simultaneously align and fold RNAs, etc. It also turned out that

the way SCFGs and HMMs are currently used may be better pictured, thanks to 2-tape MTSAGs, as the simultaneous alignment and folding of a first special tape, representing the target model, against a second tape, containing the actual sequence. This representation may lead to algorithms which will efficiently learn SCFGs from initially unaligned sequences.

While the current parser generator for MTSAGs is a usable proof of concept, which nevertheless required several months of work, I am quite convinced that there should be better ways than the current algorithm to parse several tapes. There should also exist other generalizations of CFGs which may reveal themselves fruitful. Current results are only promising starting points.

The irony of the story is that HMMs and SCFGs were borrowed by biocomputing people from other fields such as signal or speech analysis. It may very well be the time for these fields to retrofit their own models with current advances in biocomputing such as MTSAGs.