Using Stochastic Context-Free Grammars to Fold, Align and Model Homologous RNA Sequences

Yasubumi Sakakibara[*] Michael Brown[†] Richard Hughey[†]
I. Saira Mian[†] Kimmen Sjolander[†]
Rebecca C. Underwood[†] and David Haussler[†]

Stochastic context-free grammars (SCFGs) are applied to the problems of folding, aligning and modeling families of homologous RNA sequences. SCFGs capture the sequences' common primary and secondary structure and generalize the hidden Markov models (HMMs) used in related work on protein and DNA. The novel aspect of this work is that SCFG parameters are learned automatically from unaligned, unfolded training sequences. A generalization of the HMM forward-backward algorithm is introduced to do this. The new algorithm, Tree-grammar EM, based on tree grammars and faster than the previously proposed SCFG inside-outside training algorithm, produced a model that we tested on the transfer RNA (tRNA) family. Results show that after having been trained on as few as 20 tRNA sequences from only two tRNA subfamilies (mitochondrial and cytoplasmic), the model can discern general tRNA from similar-length RNA sequences of other kinds, can find secondary structure of new tRNA sequences, and can produce multiple alignments of large sets of tRNA sequences. Our results suggest potential improvements in the alignments of the D- and T-domains in some mitochdondrial tRNAs that cannot be fit into the canonical secondary structure.

[*]Fujitsu
[†]Computer and Information Sciences University of California Santa Cruz, CA 95064
haussler@cse.ucsc.edu