

Combining Stochastic Grammars

Stochastic Grammars have become a tool of wide use in Bioinformatics and Computational Biology. Originating in Noam Chomsky famous linguistic hierarchy from the late 50s, deterministic grammars were originally used in linguistic theory and complexity theory in theoretical computer science. A grammar is a finite set of rules that potentially could generate an infinite number of strings. Strings that can be generated by the grammar are said to belong to the language of the grammar. So a string either belongs to the language or not. Stochastic analogues of the deterministic grammars would assign probabilities to individual rules and indirectly strings would be given probabilities. Since the early 90s these stochastic grammars have found widespread use in describing structures in biological strings, especially the structure of genes and RNA secondary structures. These two examples exemplify the two lowest (least powerful) levels in the hierarchy; Regular Grammars for genes (corresponding to Hidden Markov Models when stochastic) and Context Free Grammars for RNA structures (corresponding to SCFG when stochastic). A series of such grammars have been devised of increasing complexity and are used to predict genes structures and RNA secondary structure. Such a grammar will take a string and annotate it in terms of gene or RNA structure dependent on the grammar, but not both. Combining grammars could be both useful and interesting. It is interesting, since it has not been tried before and will lead to new questions like how does one naturally combine 2 grammars and how could one test if the 2 original grammars were independent? It is useful, since for instance the human genome has both RNA and gene structures in it and they could overlap. The present approach would be first to search for genes and then for RNA structures and this is probably acceptable in many circumstances.

The proposed project would be to take a very simple HMM and a very simple SCFG and investigate how to combine them. A very simple HMM could jump between two states, that could correspond to coding/non-coding. A very simple SCFG could generate series of hidden states describing palindromes separated by spacers. These could be interpreted as RNA base pairing. They could be combined creating two series of hidden states simultaneously, by for instance avoiding palindromes in coding regions. Implementing these three models and investigate how easy it is to detect departures from independence.

This could be simultaneously with a literature study to investigate if there are publicly available data, where such combined grammars could eventually be applied.

Given the enormous and still growing importance of comparative genomics and that the present practice for genome annotation is to assume (with explicit statement) independence, it is surprising that this problem has not been addressed.

References.

Chomsky, N. (1956) "Three Models for the description of language" IRE Transactions on Information Theory 3(2). 113-24.

Durbin et al. (1998) "Biological Sequence Comparison" CUP

Knudsen, B. and J.J.Hein (1999) "Using stochastic context free grammars and molecular evolution to predict RNA secondary structure (Bioinformatics vol 15.5 15.6.446-454)

Pedersen, J.S. and J.J. Hein (2003) "Gene finding with as hidden Markov model of genome structure and evolution" Bioinformatics 19.2.219-227.

Pedersen, JS, IM Meyer, R Forsberg, P. Simmonds and JJ Hein (2004) "A comparative method for predicting and folding RNA structures within protein coding regions" Nucleic Acids Res. 2004 Sep 24;32(16):4925-3.