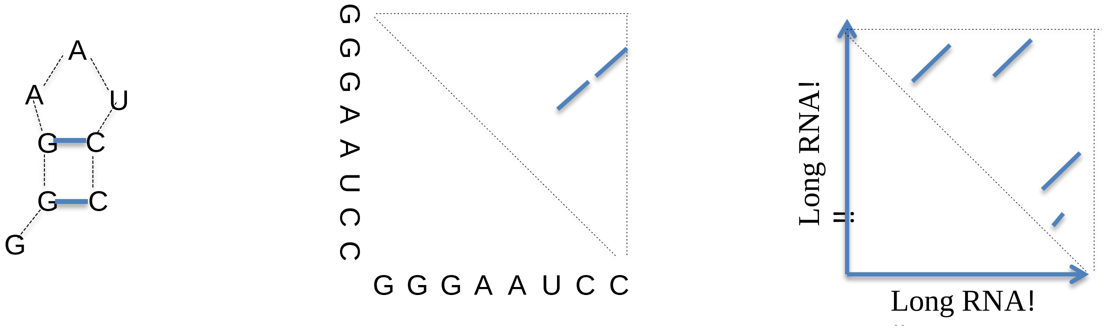
Fast stem finder

Jakob Skou Pedersen & Jotun Hein 26.8.11

Background: A hairpin is the simplest form of RNA secondary structure. It consists of a stem, potentially interupted by bulges and internal loops, and an exterior loop (see e.g. [1] for definitions of structure termninology). More complicated RNA secondary structure can be thought of as composed of hairpins, multi-loops, and what could be called inner-stems, which are not connected to an exterior loop. Because hairpins and inner-stems are fundamental units of RNA structure it is of interest to efficiently identify these.

Algorithms for predicting general RNA secondary structures (excluding pseudo-knots) run in time cubic in the length (n) of the sequence (O(n^3)). This is true both for the energy minimization algorithms and for the probabilistic stochastic context-free grammar (SCFG) approaches, which we focus on here. By restricting the prediction problem to non-bifurcating RNA structures, i.e. hairpins, the time-complexity can become quadratic in sequence length (O(n^2)). This can for instance be seen from the variant of the CYK algorithm defined for covariance models (see [1], chapter 10), which is defined for a restricted set of seven types of production rules. When the bifurcation rule is not used, the algorithm only iterates over two indices. Though writing up grammars using this restricted set of rules is not always the most compact and efficient approach, it may be convenient in this case as it simplifies the structure of the parsing algorithms. Pedersen et al ([3]) shows how the general RNA secondary structure describing grammar of [2] can be rewritten using the above mentioned restricted set of production rules.



Left: a secondary structure over 8 nucleotides, Middle: alternative representation of the same structure-sequence. Right: It is possible to find repeats and “inverted” repeats very fast. A large number of such repeats are likely to be part of the true secondary structure.

Goals: Define an SCFG describing RNA stems including bulges and internal loops that allows quadratic time parsing and hence prediction. Write up the modified versions of the CYK, inside, and outside algorithms that would be used to parse this grammar.

Perspective: The above grammar could be used to predict hairpins in RNA sequences. the parsing algorithms could be modified to efficiently identify local hairpins. A modification of the parsing algorithm would also allow the grammar to be used to predict stems between non-contiguous sequence.

However, for the the applicaition of genomic identification of functional RNA structures, it will useful to extend the grammar to handle multiple alignments instead of single sequences, as done in [2] and [3]. The grammar should then be extended with rules that describe non-structured regions to identify structured regions in seas on non-structured / non-transcribed sequence (see the grammar used in [3]). Faster algorithms exists that can find repeats in nlog(n) time and it could also be worth considering “almost perfect repeats”.

References:

[1] Durbin et al. (1998) Biological Sequence Analalysis. Cambridge University Press.

[2] Knudsen and Hein (1999) Secondary Structure Prediction Using stochastic context-free grammars and evolutionary history. Bioinformatics.

[3] Pedersen et al. (2006) Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. PLoS Comp Bio.

[4] Sean R Eddy "How do RNA folding algorithms work?" *Nature Biotechnology* **22**, 1457 - 1458 (2004)

Further comments: Jan Gorodkin also reduced complexity, but for multiple sequences. Bjarne Knudsen and Zsuzsanna Sükösd pointed us to these references:

Ogurtsov et al. (2005) Analysis of internal loops within the RNA secondary structure in almost quadratic time bioinf 22.11 1317-1324
Yair Horesh (2009) RNAslider: a faster engine for consecutive windows folding and its application to the analysis of genomic folding asymmetry *BMC Bioinformatics*, 10:76doi:10.1186/1471-2105-10-76