Poster I-103

Screening the human UTR database for stabilized RNA secondary structures

Authors:
Sven Siebert (*Department of Bioinformatics, Institute of Computer Science, Albert-Ludwigs University Freiburg*)
Rolf Backofen (*Department of Bioinformatics, Institute of Computer Science, Albert-Ludwigs University Freiburg*)

Short Abstract: We screened the human 5'UTR and 3'UTR database for stabilized RNA structures. A window sliding approach is used to fold sequence regions of various lengths. The stability of a folded RNA sequence is determined by introducing a significant value. We compared our stable regions with sequence patterns given in the database UTRSite.

Long Abstract:

Background: RNAs are known to act not only as a mediator carrying the information from the DNA to the translational machinery, but also perform catalytic and regulatory function such as the repression of translation of ferritin mRNA in the presence of an IRE element (iron response element) or the incorporation of the 21st amino acid Selenocysteine in response to a UGA codon and the presence of a SECIS element (selenocysteine insertion sequence) that appears in the vicinity of this UGA codon. These motifs are characterized through single nucleotide sequences which fold back to form Watson-Crick base-pairs A-U and C-G and the almost thermodynamically favored base-pair G-U. A base is involved in almost one base-pair. The collection of all base-pairs is called the secondary structure. To date, the two mentioned example patterns (IRE,SECIS) are well-known and can be described by motif patterns. Although they differ slightly from specie to specie, they can be trivially found in genomic sequences by pattern search programs (e.g. PatSearch). In particular, these functional patterns are found in untranslated regions (3'UTR and 5'UTR). The drawback of such search programs is to identify sequence regions based on an already known motif. These programs allow to find only sequence/structure patterns whose identification level is high, i.e. these patterns were constructed from a collection of experimentally confirmed sequences. What is more important is the fact that, in general, these pattern search programs are not capable of predicting stable structures, i.e. a found pattern does not guarantee to fulfill the assigned function due to its stability conditions. Furthermore, other sequence regions which are characterized by structural stability are excluded. A listing of stable RNA sequences in conjunction with their structures is desired.

Objectives: We propose a method that go beyond the sequence pattern descriptors and describe a screening method that systematically fold local sequence regions into their minimum free energy (mfe) structures. We therefore look for stabilized RNA regions. The stability of a folded RNA sequence is determined by introducing a significant value, that distinguishes stable regions from non-stable regions. Non-stable regions are defined to have nearly the same stability conditions as every arbitrary sequence of the same length has. A window sliding approach with size lengths ranged from 150 to 10 from each nucleotide position is used here. The folded subsequence is assigned the significance value by means

of the folding program (e.g. Mfold or RNAfold) in order to extract the minimum motif needed to describe the stable sequence region together with its structure. The mfe structure does not account for the correct structure in either case, but its energy gives clues to detect these stable regions. In doubt, one has to investigate these structures by suboptimal structures.

Results: We screened the human 5'UTR and 3'UTR database for stabilized RNA structures. A window sliding approach is used to fold sequence regions from each nucleotide position p1 of length 150 (i.e. to position p1+150) to the length 10 (i.e. to position p1+10). The computation takes several days to scan the database on a single CPU. The analytic computation time is $O(nL^3)$, where the summed length of the database sequences is n and the window size, i.e. the folding region, is L. Despite its time-consuming computation, the screening method guarantees to correctly identify the stabilized regions. These stabilized regions are compared to the already known functional patterns as given by the database UTRSite. The UTRSite database contains non-coding RNA elements such as IRE, SECIS etc. which are found by a pattern search program regardless of the stability conditions. The stability of a partial RNA sequence might be the reason why not all patterns in UTRSite could be detected by our program, thus supporting the evidence that not all patterns are functional. In addition, lots of stable sequential parts are detected, most of them which have not yet been assigned a function. Illustrations of these stable regions surveys all partial secondary structures which are significantly stable.