Poster I-23

INFO-RNAseq - Fast Inverse RNA Folding Satisfying Sequence Constraints

Authors:
Anke Busch (*Albert-Ludwigs-University Freiburg, Inst. of Computer Science, Chair for Bioinformatics*)
Rolf Backofen (*Albert-Ludwigs-University Freiburg, Inst. of Computer Science, Chair for Bioinformatics*)

Short Abstract: INFO-RNAseq is a fast algorithm for the INverse FOlding of RNA that includes sequence constraints. It designs RNA sequences that fold into a given structure and additionally satisfy some constraints on the primary sequence. INFO-RNAseq benefits from the design of a good initializing sequence by dynamic programming and the use of different local search methods.

Long Abstract:
The function of RNA molecules often depends on both the primary sequence and the secondary structure. Since prediction or experimental determination of 3D RNA structures remains difficult, much work focuses on predicting the secondary structure, which is denoted as the RNA folding problem. Here, we consider the inverse RNA folding problem, i.e. designing RNA sequences that fold into a desired structure. Additionally, we can deal with constraints on the primary sequence. Given a set of base pairs and some conditions on the sequence, we aim at an RNA sequence that is going to adopt these pairs and to satisfy the sequence constraints. Computationally approaches for predicting the secondary structure of an RNA sequence are based on a thermodynamic model that gives a free energy value for each secondary structure. The structure with the lowest energy value (called the minimum free energy (mfe) structure) is expected to be the most stable one. The general problem of the inverse RNA folding can be described as follows: Find an appropriate RNA sequence S that folds into a given secondary structure T. Thus, we have to analyze a search space of an exponentially high number of valid RNA sequences. It is hard to imagine a deterministic algorithm that finds a globally optimal solution. Therefore, local search methods are a widely used technique to address the inverse folding problem. Here, we present a new algorithm for the INverse FOlding of RNA including sequence constraints, called INFO-RNAseq, that designs an RNA sequence S that folds into a given secondary structure T and fulfills a set of given constraints on the sequence. INFO-RNAseq benefits from the design of a good initializing sequence by a new dynamic programming procedure and the use of different local search strategies. Since the success of a local search method highly depends on the quality of its initialization, an excellent initializing sequence is designed during the first step of INFO-RNAseq. This design is done via a dynamic programming algorithm similar to the RNA folding algorithm introduced by Zuker and Stiegler. It starts at the sequence design for small substructures and enlarges them gradually by one base pair and a possibly enclosed loop. Finally, we get a sequence with the lowest energy a (valid) sequence can have when folding into the target structure T. There is no other sequence satisfying all constraints that has lower energy when folding into this structure. Nevertheless, the designed sequence is not guaranteed to fold into the target structure T since actually this sequence can have even less

energy when folding into another structure. Therefore, the resulting sequence is processed further. During the second step of INFO-RNAseq, better sequences concerning the foldability into the target structure are found by doing local mutations iteratively. This is done via an adaptive walk or via a stochastic local search. To this end, a neighborhood between sequences has to be defined. neighbors to a sequence S are all other sequences S^i that differ from S in only one unbound position or in two positions that have to form a base pair concerning the desired structure T. Here, the order in which the neighbors of a sequence are tested can be chosen depending on the arising energy difference: Let E(S,T) be the energy of sequence S when folding into structure T and let S' be a neighboring sequence of S with one changed free base or base pair. Then, the energy difference E_d is given by E(S,T) - E(S',T). The higher the difference E_d, the earlier this neighbor is examined. After processing both steps of INFO-RNAseq (the initialization and the local search), we get a final sequence which fulfills all sequence constraints and which either folds with a high probability into the target structure T or whose mfe structure is T or a structure T' within a small distance from T. Using our algorithm, we got better results than the algorithm from the Vienna RNA Package that uses a random start sequence and a random order of the neighbors during an adaptive walk. Furthermore, our algorithm is much faster (despite it consists of two parts) since less steps are needed during the local search. This shows that the designed sequence from the first step is an excellent initializing sequence.