Secondary Structure Prediction of Large RNAs

Michael Geis

Generally, an understanding of RNA thermodynamics and folding kinetics can be derived from its secondary structure. Experiments have shown that alternative structures of an RNA can perform different functions. Thus, to understand the functionality of RNAs, a study of its folding and re-folding behavior is mandated.

This work covers the Kinwalker and HelixPSO algorithms. Kinwalker predicts RNA folding trajectories, i.e. series of intermediate states connecting the initial structure with the predicted structure. The folding process is split into a series of interlaced folding and transcription events. The re-folding mechanism exploits that known metastable states apparently consist of energetically favorable combinations of locally optimal substructures. Kinwalker estimates the first passage times of folding events based on the energy barrier between successive structures. The energy barrier height is computed via a heuristic proposed by Morgan and Higgs as well as variations and extensions thereof.

Kinwalker's predictions show excellent qualitative agreement on a set of sequences with experimentally well-characterized folding pathways. The estimated folding times are mostly accurate. Kinwalker can compute RNAs of up to 1500 nucleotides, covering most RNAs for which kinetic effects are known to play a crucial role. Thus, Kinwalker can handle much longer sequences than other algorithms working at base pair step resolution and makes more accurate predictions.

HelixPSO is a Particle Swarm Optimizer (PSO) algorithm, a biologically inspired optimization technique imitating swarm behavior. It simulates a clustered swarm of particles collectively exploring the fitness landscape of the RNA secondary structure space. The particles share knowledge about the search space with each other. The points in the space are the possible secondary structures of the input RNA, represented by a permutation of an ordering of the set of possible helices. Particle movement is conducted by transforming one conformation into another by swapping indices. The performance of HelixPSO – measured by free energy or number of correctly predicted base pairs – is compared to a set of algorithms implementing Dynamic Programming (RNAfold), Genetic Algorithm (RnaPredict), Simulated Annealing (SARNA-Predict) as well as PSO (SetPSO) methodologies. When free energy is minimized, HelixPSO consistently achieves lower values than RnaPredict and SetPSO.

In base pair prediction, HelixPSO preforms close to RNAfold for average scores. For best values HelixPSO outperforms RNAfold by 9% in sensitivity, 18% in specificity and 13% for the F-measure. HelixPSO outperforms RnaPredict and significantly outperforms SetPSO. HelixPSO does almost as well as SARNA-Predict using the INN and INN-HB energy models. When compared to SARNA-Predict using the advanced efn2 energy model, SARNA-Predict clearly outperforms HelixPSO. For average instead of best values, HelixPSO performs significantly better than SARNA-Predict and SetPSO.

PSO is a new approach to the RNA folding problem and HelixPSO performs

very well in comparison to established algorithms such as RnaPredict, marking a significant improvement over SetPSO, the only other PSO algorithm for the problem. The performance of HelixPSO in comparison with RNAfold, which is very similar to the benchmark RNA folding algorithm mfold, demonstrates that PSO algorithms are well suited to the problem domain of RNA secondary structure prediction.