Methods

Base Pair Tree Representation of Secondary Structure

RNA secondary structure can represented by base pair trees, where only base pairs are considered. This representation is very similar to level 5 RNA abstract shape [1], albeit the stem length information is conserved. Although this representation greatly increases the size of the auxiliary space (codomain) compared to level 5 abstract shapes, it does not suffer from major changes in representation if small stem-loops are added or removed. It also does not map structures of widely varying stem length to the same element in the codomain.
The mapping from a Vienna dot bracket representation can be made simply by removing the dots (unpaired nucleotides) and then creating a tree from it. Since all nodes represent base pairs, the resulting tree can be considered unlabelled. Note that an artificial root is added to all structures.

Unit Tree Indel Distance

The unit tree indel distance is a simple and intuitive distance function based on the well known tree edit distance [2, 3], where insertions and deletions have a unit cost. Since it is applied to base pair trees which are unlabelled, node relabelling comes at no cost. Under this scoring scheme, a distance of n means that the smallest number of base pair breakage or formation to transition from a base pair tree to another is n. The distance function retains the metric properties.
Overall, the unit tree indel distance is an interesting distance function with both intuitive interpretation and very useful mathematical properties.

Minimizing Sum of Pairwise Distances

Minimizing the sum of pairwise distance between sets of objects is far from new and a significant body of work has been done in the field of location theory [4].

The consensus problem described here can be reduced to an instance of the minimum average distance placement (MIN-SUM) by modifying the distance function in use such that elements belonging to the same group have infinite distance, thereby guaranteeing that a consensus will never contain more than one element of each group. This convenient reduction allows us to benefit from previously derived hardness results and justify the use of meta-heuristics.
Using a metric distance function, one might think that it would be possible to reduce efficiently the search space, but even finding a near-optimal placement within a certain factor of the optimal has been proved to be NP-Hard [4].

Consensus Solvers

Given the complexity of the consensus problem, the use of heuristics is well justified. Many simple heuristics were found to perform well within reasonable time. These algorithms all take precomputed distance matrices whose calculation is embarrassingly parallel.

Exact Version: Branch & Bound First, an exact version of the consensus solver based on a branch & bound design was created to compare the performance of different heuristics on small instances.

The lower bound of an incomplete solution is estimated by the summation of pairwise distances between the chosen objects and the best distance they have to remaining sets of objects to be chosen. This bound is generally quite good on small instances, but its performance degrades rapidly as the number of molecules grows.

Heuristic: Genetic Algorithm The genetic algorithm solver uses generic mutation (random substitution) and crossover operators (uniform) along with elitist selection. To increase convergence rate, an improvement operator using a steepest descent procedure is used.
The local search is done by finding the substitution which improves the most the current sum of pairs distance. The operator applies a small number of iterations of local search with a certain probability.

MC-Cons

MC-Cons is a two-step optimization procedure (Fig. 1). First, *base pair tree consensus* are calculated. This insures that the final consensus will minimize the difference in base pair topology. For each base pair tree consensus and for each group of structures, the subset of secondary structures whose base pair tree fits the one assigned to that group is kept. String edit distance consensus (on Vienna dot brackets) are then calculated. Normally this distance function would not be sufficient, but since tree distance is already minimized, the final consensus is satisfactory.

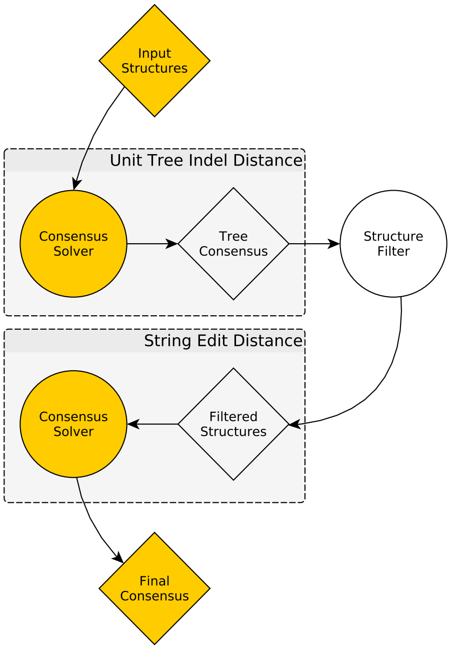

Figure 1. Flowchart illustration of MC-Cons procedure. The two-step procedure is quite simple. The user only inputs data and chooses which consensus solver, either the exact one for small instances or the genetic algorithm for larger problems.