Inferring Pseudoknotted RNA Secondary Structures from SFold Output

Aug. 28, 2008

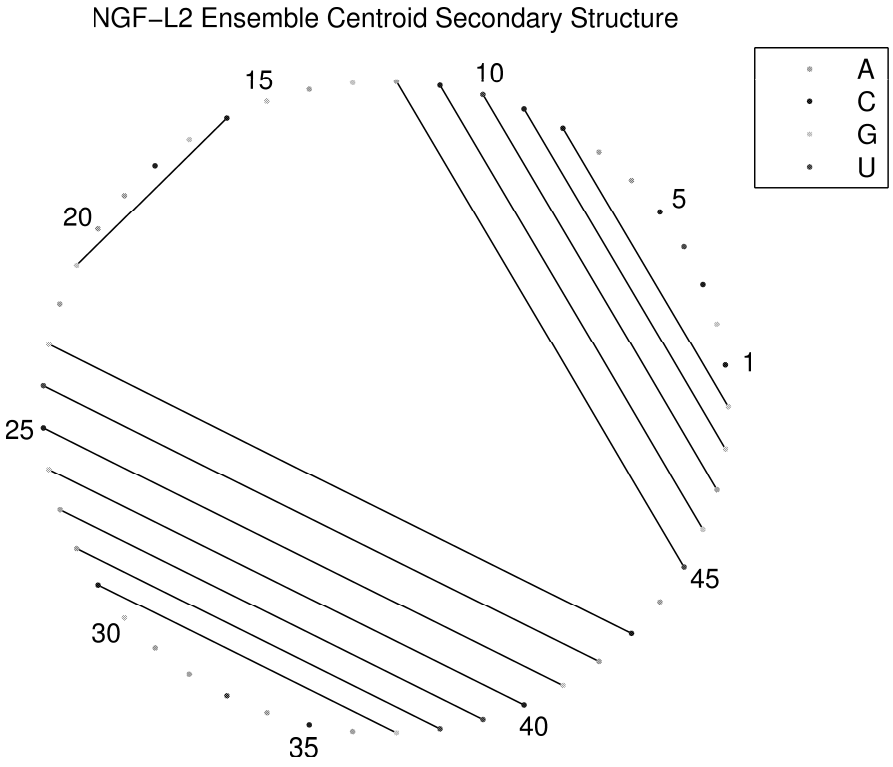
SFold is an algorithm that draws samples from the ensemble of RNA secondary structures in proportion to their Boltzmann weights (Ding et al., 2006). Thus, it guarantees the generation of a statistically representative sample of the Boltzmann weighted ensemble of structures. From the sampled set of structures, it calculates a centroid structure to characterize the central tendency for the set of structures. The centroid for a given set of structures is the structure that has the minimum distance to all of the structures in the set. The distance between two structures is the number of different base pairs. The algorithm also clusters the sampled structures and calculates a centroid for each cluster. (Ding et al., 2005) have shown that in high-dimensional discrete inferences, the most probable solution often has low probability, indicating that no single solution can represent the posterior space well. They also show that for RNA secondary structure that centroid solutions better represent the full posterior weighted ensemble of solutions, and yield more specific predictions.

Most algorithms for secondary structure prediction are based on a decomposition of the base-pairing graph for a molecule into distinct loops that are associated with thermodynamic values based on loop sequence, length and type (Dirks and Pierce, 2003). Depicted as a circle with arcs connecting paired bases, the predicted secondary structures appear as bundles of arcs with no crossing lines. Figure 1 shows the SFold predicted centroid for a 49 nucleotide section of NGF-L2, ligand L2 of human nerve growth factor (Binkley J. et al., 1995). The calculated energy for this structure is -17.50 kcal/mole.

Pseudoknots are formed when two base pairs i, j and d, e with $i < j$, fail to satisfy the nesting property $i < d < e < j$. In a circle diagram, this would be represented as crossed arcs. Pseudoknots are known to exist in ribosomal RNA, viral RNA and a number of ribosomes. Pseudobase (Batenburg et al., 2000) lists over 200 naturally occurring RNA pseudoknot structures from a variety of different organisms.

One limitation of the SFold algorithm and most other RNA secondary structure algorithms is that they do not allow structures containing pseudoknots. Most secondary structure prediction algorithms, including SFold, use dynamic programming to compute the partition function needed to evaluate the probability of folding to a particular secondary structure. If pseudoknots are excluded,

Figure 1: SFold predicted centroid for NGF-L2, ligand L2 of human nerve growth factor. The energy of this structure is -17.50 kcal/mole.



the forward step of the dynamic programming algorithm has time complexity of $O(N^3)$ and uses $O(N^2)$ memory, where N is the length of the sequence. It is possible to extend the dynamic program to include pseudoknots, but the algorithm has a time complexity of $O(N^5)$ and requires $O(N^4)$ memory. It would be possible to extend SFold to calculate the partition function and sample pseudoknot structures, but the computational complexity makes the calculation too time consuming for other than short sequences. However, given an RNA secondary structure, the energy of that structure may be calculated in $O(N^2)$ steps.

Rather than calculating the partition function for structures containing pseudoknots, we propose to use the set of SFold samples to generate additional secondary structures, possibly containing pseudoknots. To generate new structures, we sample a pair of structures from the collection of SFold samples structures. SFold typically samples 1000 structures from the posterior Boltzmann distribution. Since SFold calculates a partition function, it also calculates a Boltzmann probability of the structure. Each structure consists of a list of base pairs, along with the corresponding energy and probability. Pairs of structures are combined into a single list of base pairs. Duplicate base pairs are removed. In the case where a there is a single base on the list that is paired with two different bases, a fair coin is flipped to decide which pair to include in the new structure. Combining the structures in this way can produce pseudoknotted structures.

We sample new structures in this manner to generate proposal from the distribution of RNA secondary structures, including those with pseudoknots. We use a Metropolis-Hastings (MH) algorithm to generate a sequence of samples from the distribution of secondary structures.

The algorithm

1. Sample a pair of structures (uniformly), a, b, from the SFold sampled structures.

2. Combine the structures into a new structure, *current*. Let $t_{current}$ be the number of coin flips required to break ties.

3. Set $sample - list = \{\}$.

4. Repeat steps 5 through 7 for a fixed burn-in period, followed by a fixed sampling period.

5. Sample a new pair of structures, i, j. Combine the structures into new structure, *proposal*. $t_{proposal}$is the number of coin flips required to break ties.

6. Calculate the transition functions $T(proposal|current) = T(propsal) = P(S_i)P(S_j)(0.5)^{t_{proposal}}$ and $T(current|proposal) = T(current) = P(S_a)P(S_b)(0.5)^{t_{current}}$. The proposal does not depend on the current structure pair.

7. With probability $\min(1, \frac{exp(\frac{-energy_{proposal}}{RT})T(current)}{exp(\frac{-energy_{current}}{RT})T(proposal)})$ set $current = proposal$; $i = a$; $j = b$. R is the universal gas constant and T is the temperature, 37° Celsius. If the burn-in iterations have been completed, add $current$to the $sample - list$.

8. Calculate a centroid structure from $sample - list$by combining all base pairs with a sampling frequency greater than 0.5.

Structure energies are calculated using the *Nupack* program. In some cases, *Nupack* is unable to calculate the energy for a given structure. These structures are assigned a large positive energy value.

A problem that arises with this algorithm is that proposal acceptance rate is very low, on the order of 1%. This indicates that the algorithm is not fully exploring the solution space.

NGF-L2 Example

For the sequence in Figure 1, SFold clustered the 1000 NGF-L2 sampled structures into two clusters. The first containing 79% of the samples and the second containing the remaining 21%. The centroids for these clusters are shown in Figures 2 and 3.

The pseudoknot algorithm described above predicted a centroid structure containing a pseudoknot. The predicted structure combines the major features of the two cluster centroids. Figure 4 shows the predicted structure. This structure has a calculated free energy of -22.3 kcal/mole. This is lower than the MFE non-pseudoknot structure.

The predicted structure matches the NGF-L2 structure downloaded from Pseudobase (http://wwwbio.leidenuniv.nl/~Batenburg/PKB.html) shown in Figure 5.

.

Figure 2: Cluster 1 Centroid

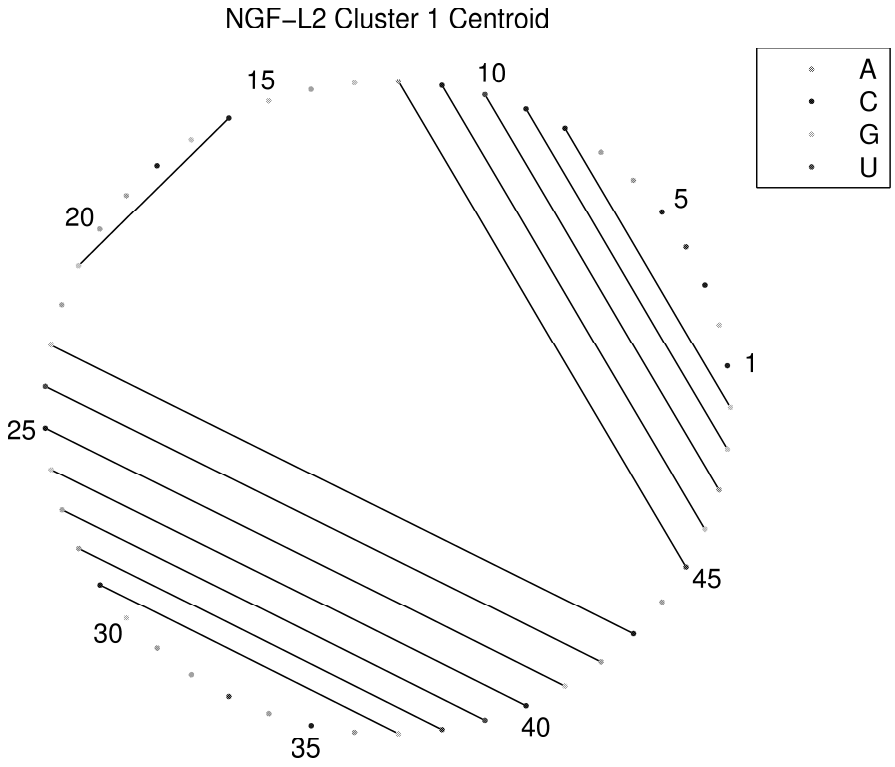


Figure 3: Cluster 2 Centroid

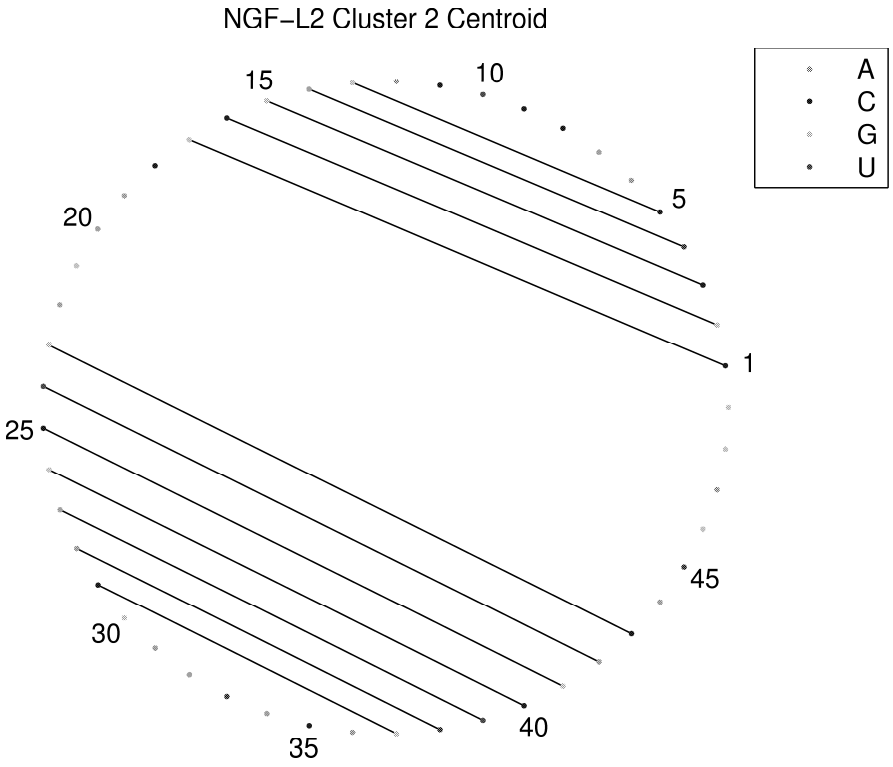


Figure 4: Predicted Structure

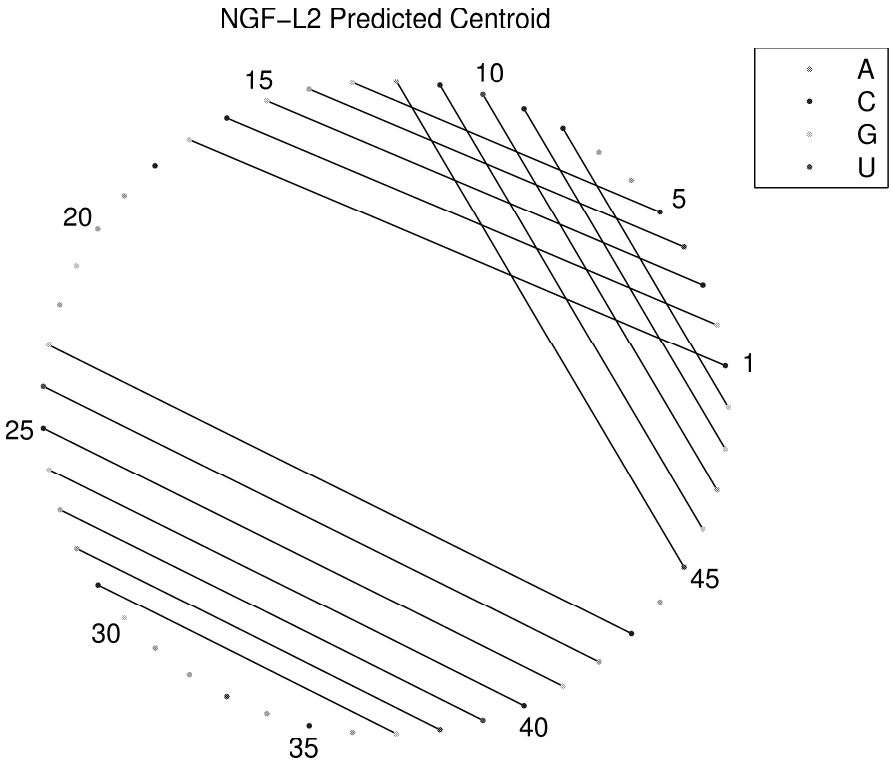


Figure 5: NGF-L2 Structure from Pseudobase