Pseudoknot prevalence and base pair identity distance in prediction of secondary structures

The distance $D^{\zeta,\zeta'}$ between thermodynamic ensembles of secondary structures ζ including pseudoknots and ζ' excluding them is defined as:

$$D^{\zeta,\zeta'} = \frac{\Sigma_{i,j}|P^{\zeta}_{i,j} - P^{\zeta'}_{i,j}|}{\frac{1}{2}\Sigma_{i,j}(P^{\zeta}_{i,j} + P^{\zeta'}_{i,j} + |P^{\zeta}_{i,j} - P^{\zeta'}_{i,j}|)}, \text{ where } P^{\zeta}_{i,j} = \sum_{k=0}^{n} \frac{e^{-\beta E_k}}{Z_\zeta}\delta^{\zeta}_{k}(i,j) \text{ and } Z_\zeta = \sum_{k=0}^{n} e^{-\beta E_k}$$

and $\delta^{\zeta}_{k}(i,j) = 1$ iff i and j are paired in the kth structure of the thermodynamic ensemble ζ and $\delta^{\zeta}_{k}(i,j) = 0$ otherwise. With this definition, the distance distribution between the two thermodynamic ensembles of secondary structures (right graphs) would be exactly equivalent to the distribution of pseudoknot proportion (left graphs[19]) *if* pseudoknots could just be added to secondary structure predictions excluded them. The fact that right graphs are virtually flat (except for sequences having essentially no pseudoknots) reveals the strong cooperativity of secondary structure rearrangements. Statistics is done on thousands of random RNA sequences[19]. Number of lowest free energy structures included in each thermodynamic ensemble: n=3.

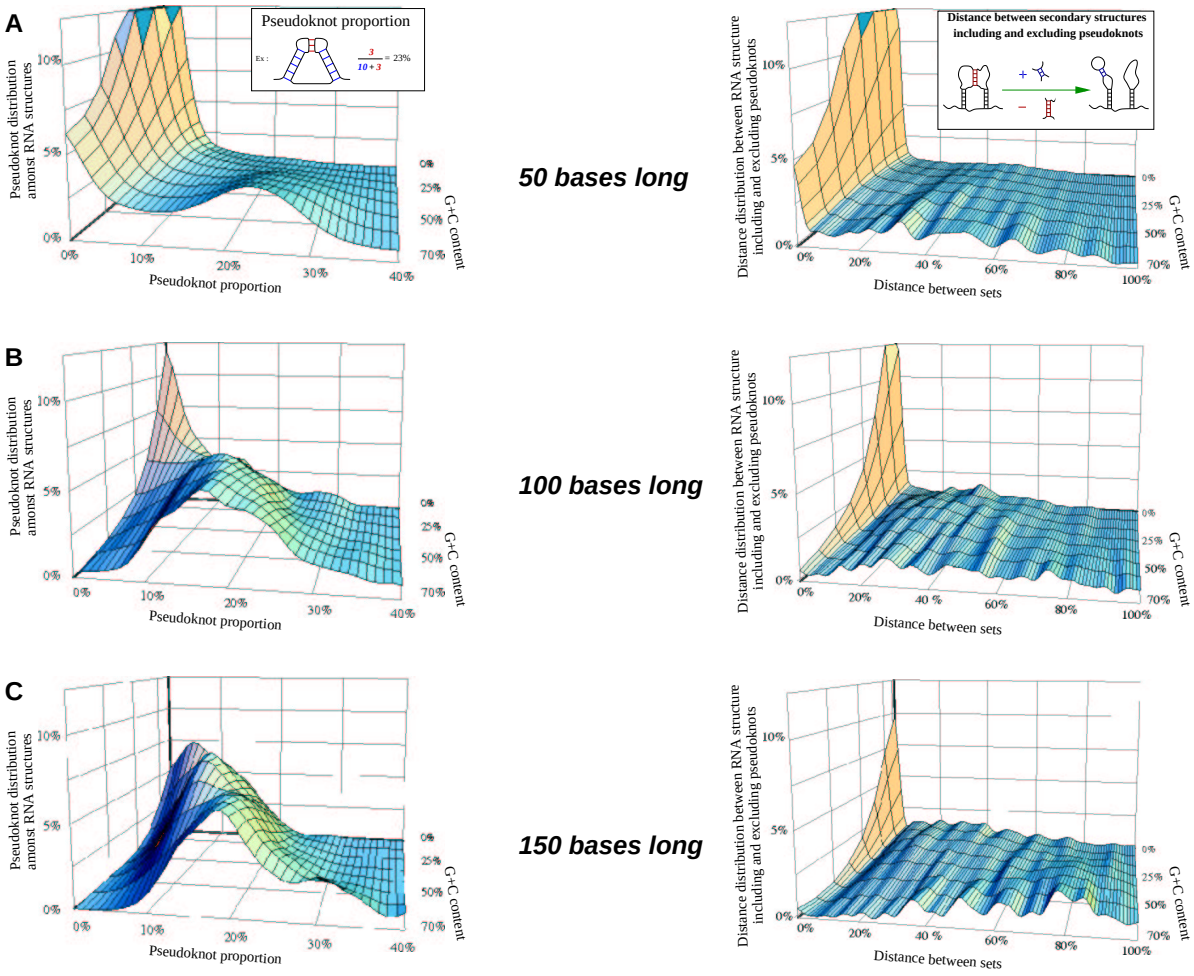

Figure 4. **Right:** distribution of pseudoknot proportion amongst formed base pairs for 50-nt-long (A), 100-nt-long (B), 150-nt-long (C) random RNA sequences of increasing G+C content. **Left:** distribution of distance between low energy RNA secondary structures folded with and wihtout pseudoknots.