RNA sequence analysis using covariance models
Notes by Elizabeth Tseng

1. CM structure: hair pin loops usually have at least 3 nucleotides (2 is unlikely because the molecule can't bend that sharply), and 4 is quite common.

2. Model training: the difficulty in model construction is deciding which are pairing states, which are single and which are bifurcations.
   ➔ Solution: mutual information (MI)

3. Mutual Information (MI): highly conserved nucleotides in aligned sequences could imply possible pairing. If one position is perfectly conserved, then M.I. = 0.

4. M.I. example (see slides): column 6 & 7 have no biological relation but there is some mathematical implication – if the nucleotide in column 7 is U, then the corresponding nucleotide in column 6 is limited down to A & U, and the same for G, C, A in column 7. This yields M.I. = 1. If we know something about the nucleotide at column 7, then we can well predict column 6.

5. M.I. is most seen at Watson-Crick pairs, which suggests that they're correlated.

6. The M.I.-based structure learning is comparable to Nussinov/Zucker except that it uses M.I.. For good structure learning, we need sequences of appropriate phylogenetic distances; if the sequences are too conserved, then we get no mutual information. If they're too diverged, tey're very hard to align.

7. The building of Rfam: experts hand-align a particular RNA family to produce a "seed alignment" and secondary structure consensus. Then they scan through a small fragment of the database to fine-tune parameters (threshold, window length, etc.).