CSE 527, Lecture 16, 11/20/2006 Noah Benson

RNA Pairing and Secondary Structure Prediction

- Can maximize number of base pairs or minimize energy
- Alternately, look at loops – better for modeling
- Note: CG pairs have 3 H-bonds while AT pairs have 2, thus CG pairs are more stable.
- Zuker algorithm – loop based
  - dynamic programming technique based on two matrices W and V; $V(i, j)$ represents energy if i is forced to pair with j.
  - V chooses minimum energy based on different kinds of loops – whether i, j are part of a hairpin or stack, close a bulge, or are in a multi-loop.
  - Ultimately requires empirical information about the energy of the various interactions; these come from experiments.
  - $O(n^4)$, but $O(n^3)$ under (reasonable) simplifying assumptions.
- Ultimately interested in tertiary structure, but these algorithms can only compute secondary structure (tertiary is still too hard)
- We might be interested in suboptimal energies because the conformation of a molecule is not necessarily always the lowest energy – a variation in temperature, for example, can cause shifts in RNA structure when the energy between states is small
- Can modify Zuker's algorithm to find suboptimal folds
- McCaskill algorithm finds a probability distribution over all the energy states
- Alternate structures are important (at least) for some switch-like control functions of RNA.
- Estimates suggest reasonably high confidence (50%-75%) in relatively small molecules (300nt or fewer)

RNA Motif Description

- Use HMMs – given a multiple alignment, we can train it (using Baum-Welch) to have probabilities of outputs (nucleotides) at positions. Given a new sequence, we estimate the probability that the sequence belongs in the alighment.
- This is good for sequence, but it ignores structure
- Can get around this somewhat by pairing up columns with an additional variable that describes motifs. A row whose columns describe secondary structure in the alignment.

- This can no longer be modeled by an HMM, due to the interactions of distant column pairs, but a more elaborate model (below) can do so. It's much slower, but more accurate.
- covariance model – captures this covariation between columns. the covariation typically reflects mutations between columns that keep an RNA structure stable (eg one nucleotide mutates so the nucleotide that lies across from it/bonds with it mutates to compensate)
- this is essentially a stochastic context-free grammar
- Durbin & Eddy paper – gives accurate search for tRNA – still not quite as good as human experts but close.
- Probabilistic model – compare the probability according to the model to probability of a background model.
- Like HMM – can use forward algorithm or Viterbi algorithm. Can also be used for alignment/structure prediction
- in comparison between RNAs, RNAs that were believed to be tRNAs score highly while RNAs believed not to be tRNAs scored low. Very important to have good separation because there are billions of nucleotides and only a few tRNAs.
- High alignment quality – close to experts (better than ClustalW)
- tRNAscanSE – current favorite CM-based RNA scan. Speeds up the CM algorithm using older heuristic algorithms
- Covariance Model
  - contains several parts:
  - sequence with a structure
  - "guide tree" – each node is like a state – emits twice (CGUA or - each with some probability) following an in-order traversal. Doesn't give structure information explicitly but can be inferred from the tree.
- CM can use a Viterbi-like algorithm
  - Recursively decides on the bset choice for state change based on the best way to produce the rest of the string given that change.
  - Can use dynamic programming effectively – $O(qn^3)$ where q is the number of states. This is because the bifurcation causes the n^3 instead of n^2 – we must decide where to bifurcate within a string of size $\leq n$.
  - Also used in NLP and speech processing
- Covariance Models can be trained as well