CSE 527 Larry Jean
11/29/06 leijean@amath.washington.edu
Lecture 19

Paper: Distance vs Accuracy
- low distance → low accuracy
- greater distance → greater accuracy
- increase distance more → decrease accuracy, levels out
- structure aligning adhoc, useful but not tremendous

CM Finder
- simultaneous aligns and predicts structure
- idea(heuristic):
  - pick out interesting regions to start
  - *EM* iteration
  - realign (via *Viterbi*)
- use mutual information + folding energy to predict structure
- Heuristics: "finding candidate"
  - scan sequence & look for low energy for candidate
  - *tree edit* ←→ Vienna algorithm
    - how to convert one tree to another
    - secondary structure of RNA can be abstracted to tree, not much evolutionary considerations
    - look for similarities
    - can calculate closest of all previous candidates (minimizes sum of distances to all others)
    - result: generate a sequence of candidates
  - for every candidate set → apply *EM*
  - can have different & strong *BLAST* sequence match
    - align *BLAST* anchors to the candidate sequence
- How to build structure model?
  - got alignment, maximize joint probability of data & structure
  - assume independence of unpaired columns
  - within column pair model dependence
  - no prior knowledge of what's paired:
    - $I_{ij} = \log (P(L_iL_j)/P(L_i)P(L_j))$ → sum of mutual information terms
  - have prior knowledge of what's paired:
    - $P(D,\sigma) = P(D \mid \sigma) * P(\sigma)$
      $= (\text{single stranded product})*(\text{double stranded product} = K_{ij})$
    - D = data, sigma = structure
    - $K_{ij} = I_{ij} + \log(P_{ij}/(s_i*s_j))$ ← prior information
    - Question: how to know of prior information?
      - take single structure estimate and thermodynamics
      - not rigorously "prior" in Baysian sense, but heuristically has the same effect

- *CM Finder* works best on Rfam families w/flanking sequences versus *RNA Alifold*, *CARNAC*, *FOLDALIGN*
- Table:
  - sequence length range widely
  - *CARNAC* has high specificity but low sensitivity (tradeoff)
  - *CM Finder* has better balance

Applications of CM Finder

- look for RNA elements in prokaryotes
  - goal: infer structure prediction of these RNA
  - more efficient to search for cis-regulatory RNA elements
  - use comparisons between genome
- Approach:
  - pick favorite bacteria
  - find close othologous (BLAST/CDD)
  - best genes (Footprint finds patterns)
  - *CM Finder* for structure motif
  - search genome database for more homologs to narrow down candidates
- *Footprinter*:
  - find small patches that are nearly identical from one sequence to next
  - suppose to allow no gap, but gaps interesting because might be hairpin, etc.
  - test successful interesting patterns (turns out to be T-box in this case)
  - amino acid and t-RNA joined by amino acyl tRNA-synthetase:
    - tyrS effects uncharged tRNA
    - yes/no amino acid attached effects its shape
    - if uncharged, causes downstream genes to produce more tRNA-synthetase to charge it
- Results:
  - Want to rediscover things that are known, to reinforce novel results
  - Ranking of Rfam family
    - Specificity low → mixture of two groups and only found half
    - 30~40% of bacterial energy goes to ribosomal protein → how to coordinate?
    - Ex.: when L19 bound/unbound, different shapes of mRNA leader for that ribosomal protein

Future Works

- Better identifying duplicates, improve rankings
- Scale up to eukaryotes, but 2~3 orders of magnitude more work to do that
- Summary:
  - *Covariance*: powerful, expensive
  - *Rigorous/Heuristic filtering*: faster, low loss in accuracy
  - *CM finder*: CM based motif discovery