CSE527 Lecture Notes 11/28

Adrienne Wang

1 Review: RNA Secondary Structure

Nussinov = MAX pairing
$B(i,j) = \#$ of pairs in optimal pairing of $r_1...r_j$
$B(i,j) = 0$, $\forall i,j$ with $i \geq j-4$
case 1: $B(i+1,j)$, i doesn't pair
case 2: $B(i,j-1)$, j doesn't pair
case 3: $B(i+1,j-1)+$ if r_i pairs with r_j then 1, else 0, neither pairs
case 4. $max\{B(i,k)+B(k+1,j)|i<k<j\}$
Time complexity for this algorithm is $O(n^3)$. n^2 entries in the matrix and n choices for k. This algorithm may be redundant in some sense, but covers all cases.

2 RNA Sequence Analysis Using Covariance Models by Eddy & Durbin

Probabilistic model for RNA families: It's known as a Covariance Model, which is analogous to Stochastic CFG, and also generalization of a profile HMM.
Training can be performed from aligned/unaligned sequences (standard alignment algorithms do not consider secondary structure). It automates "comparative analysis", and complementes Nusinov/Zucker RNA folding
Ex:
A C G G A U C G
A C – G A U – G
A G A G A – T C
Alignment tools might try to push the first G in the third sequence to next

position.

3 Main Results

It's a very accurate search for tRNA (before the current favorite tRNAscanSE). It models construction comparable to human experts given sufficient data. It gives quantitative info on pseudoknots and other tertiary features.

4 Probabilistic Model Search

Given a sequence, calculate likelihood ratio that model can generate the sequence vs. a background model. Set a score threshold, above the threshold means a hit.
Scoring:
forward/inside algorithm - sum over all paths
Viterbi approximation - find single best path (Bonus: alignment & structure prediction)
Ex: search for tRNAs (see lecture slides for the graph)
good separation for real tRNAs.

5 Alignment Quality

U100: slight mismatch, but get big components right.
ClustalV: bad misalignments for RNA sequences. (ClustalV is a tool which considers sequence only)

6 Comparison to TRNASCAN

Fichant & Burks: best heuristics. 97.5% true positives and 0.37 false positives per MB.
CM A1415 (trained on trusted alignment): $>$ 99.98% true positives, $<$ 0.2 false positives per MB
tRNAscanSE (current favorite): CM-based with heuristic pre-filtering to

make it faster. The parameters of the pre-filters are tuned for generous results.

7 CM structure and Overall CM architecture (see figures in lecture slides)

CM models a sequence as a "guidetree", representing what's paired and what isn't. Each branch is an HMM emitting both sides of a helix.
Architecture: a generalization of HMM. Each box - a node of guide tree.
Left singlet node and right singlet node are just like HMMs. 3 possible states (MATL(MATR for right singlet node), INSL (INSR for right singlet node), DEL) represent the ways that a letter could be emitted from this node.
MATP emits pair of symbols (16 possible pairs), modeling base pairs in RNA secondary structure.
BIF represents a junction between multiple helices in the secondary structure. It allows multiple helices.

8 CM Training

Given a set of unaligned sequences, the training algorithm gets a random alignment, and starts to estimate parameters of the CM based on that multiple alignment. The set of sequences is aligned again based on the constructed model, and standard EM iterations are repeated until convergence. The alignment is done through Viterbi algorithm.

9 Mutual Information

MI is the expected score gain from using a pair state. It calculates the amount of information gain between 2 columns in alignment.
$M_{ij} = \Sigma_{x_i,x_j} f_{x_i,x_j} log_2 \frac{f_{x_i x_j}}{f_{x_i} f_{x_j}}$, $0 \leq M_{ij} \leq 2$
f_{x_i} is the frequency of seeing x in the ith column
M_{ij} is max when no sequence conservation but perfect pairing.
if $M_{ij} = 0$, 2 columns are independent
if $M_{ij} = 2$, the first column perfectly predicts what's in the 2nd column.
It's a NP-hard problem finding optimal MI. If we need to find optimal MI

without the pseudoknots, dynamic programming can accelerate the running time.