CSE 527 Notes: Lecture 19, December 1, 2004
Robyn Greaby

Covariance Models Cont.

Alignment
Verterbi:

- 'Inside Algorithm' – analogous to 'Forward algorithm' for HMMs
- Inside: find the most probable sequence of transitions and emissions to produce the sequence
- Forward: what is the probability that the sequence was produced by the model

Viterbi Algorithm for covariance models

$S_{ij}^{y} \cong \log \Pr$ {Substring i, i+1, …, j generated started at state y}

T = transition probability
E = emission probability

(1) If y is a Match Pair State

$$S_{ij}^{y} = \max_{z}\left[S_{i+1,j-1}^{z} + \log T_{yz} + \log E_{x_i x_j}^{y}\right]$$

(2) If y is a Match/Insert left or right (like forward for HMM)

$$S_{ij}^{y} = \max_{z}\left[S_{i+1,j}^{z} + \log T_{yz} + \log E_{x_i}^{y}\right] \qquad \leftarrow \text{left}$$

$$S_{ij}^{y} = \max_{z}\left[S_{i,j-1}^{z} + \log T_{yz} + \log E_{x_j}^{y}\right] \qquad \leftarrow \text{right}$$

(3) If y is a Delete State

$$S_{ij}^{y} = \max_{z}\left[S_{ih}^{z} + \log T_{yz}\right]$$

(4) If y is a Bifurcation state

$$S_{ij}^{y} = \max_{i+1 \le k \le j}\left[S_{i,k}^{y_{left}} + S_{k+1,j}^{y_{right}}\right]$$

- Need to max over all values of k because there are insert and deletes, so can't directly predict where the split will be
- Slows down algorithm

- in HMM fill in n*states table: O(n*states)
- extra dimension in covariance models because it is necessary to deal with bifurcations: O(n^2*states)
- usually few bifurcation states but still O(n) slower

Training: Mutual information

$$M_{iu} = \sum f_{x_i x_j} \log_2 \frac{f_{x_i x_j}}{f_{x_i} f_{x_j}}$$

$0 \le M_{ij} \le 2$

-Max when no sequence correlation, but perfect pairing

- Independent = 0
- Always paired = 2
- Mutual information: Expected gain in score from using pair state (versus 2 single states)
- Model with pair states columns in optimal alignment with high mutual information
- Optimal MI: NP hard (optimal pairing of columns) (?)
- Optimal MI without pseudoknots: dynamic programming (not NP hard)

Training: Algorithm

$$S_{ij} = Max \begin{cases} S_{i+1,j} & \leftarrow \text{i is not involved in a pair} \\ S_{i,j-1} & \leftarrow \text{j is not involved in a pair} \\ S_{i+1,j-1} & \leftarrow \text{i and j are paired} \\ \max_{i<k<j} S_{i,k} + S_{k+1,j} & \leftarrow \text{i and j are both involved in pairs, but not with each other} \end{cases}$$

S = max mutual information

- Builds a n*n upper triangular matrix
- Form alignment based on mutual information
- Pair states use rule 3 in recurrence
- Once trained use Viterbi to find more RNAs

How bad is pseudoknot constraint

- Ignoring pseudoknots speeds things up
- May reduce sensitivity
- Is there a better way?

An upper bound on optimal mutual information:

$$\sum_{i=1}^{n} (\max_j M_{ij})/2$$

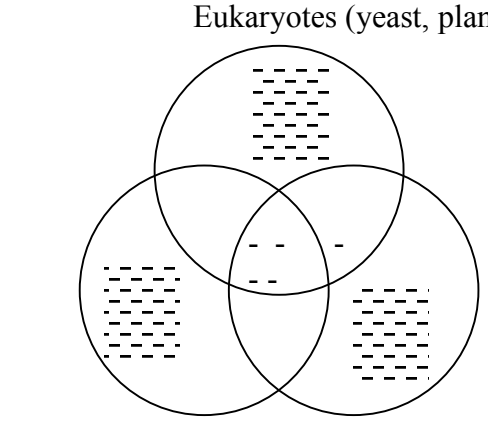
- Not necessarily realizable because multiple i's may choose same j, but only one can match with j
- Only 2 bits of additional information without pseudoknot constraint
- Lots of information in 2˚ structure, but not much in pseudoknots

Rfam: an RNA database

1/2003 release 1.0: 36 entries	entries = families of RNAs
6/2004 release 6.1: 379 entries	280,000 sequences

- Biggest scientific computing user in Europe
- 1000 cpus for a month per release
- Built on covariance models

- Takes a trusted alignment and builds covariance models then looks for more members of the family



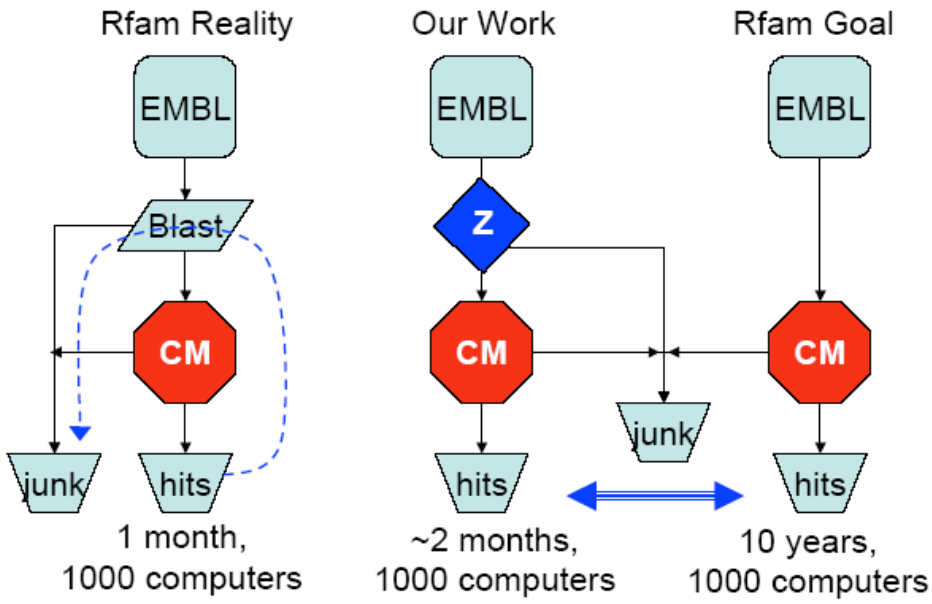
Bacteria
- Lots of data because well studied

Archaea
- Diverged from bacteria 4 billion years ago
- Live in extreme environments

- tRNAs, rRNAs and 2 other families are the only ones common to all 3 domains

Rfam: details

- Hand curated 'seed' alignments with structure annotation
- Starts with ~10 members of family (more if possible)
- Does not use mutual information (possibly because seed families are small)
- Build covariance models (use EM model)
- Search database for new family members
- Blast everything for 7 base sequence identity
- Accidentally throws out some things that are wanted



(Figure from lecture notes)
New method for filtering out uninteresting things to replace blast

- Slower
- But everything thrown out is certain to be uninteresting (below CM's score threshold)
- Works on 90% on the families in Rfam
- Found many new RNAs
- May have false positives, but all things found are real based on CM
- Many newly found RNAs appear to be sensible
- Results more sensitive than blast.