CSE 527, Notes for lecture 17, 11/22/06 Field Cady

Task 2: Motif Description
A Covariance Model (CM) accounts for folding, and sequence, rather than simply sequence as in a hidden Markov model.

CM Viterbi alignment:
S_{ij} = best way to emit from i to j
This is either

1) i and j match
2) match/insert left/right
3) delete
4) bifurcation, where a new loop is generated

Downside: O(qn^3) runtime where q is number of states, n is length of strand

Model training: good alignment and structure <--> covariance model

How to get structure? Use Mutual Information (MI). $M_{ij} = \text{sum}_{xi,xj}\{\log(f_{xi,xj}/f_{xi}f_{xj})\}$
Finding the optimal MI can be done with dynamic programming
$S_{ij} = \max\{S_{i,j-1}, \max_{i<=k<=j-4}(Si_{,k-1}+M_{k,j}+S_{k+1,j-1})\}$
We need enough sequences to do this, and at an appropriate phylogenetic distance from each other.
happily: accounting for pseudo-knots doesn't bolster this model much

tRNAscan SE=program used to identify new tRNA. Uses a prefilter to save time, and then does a CM on those that pass the filter.

Rfam=database of RNA families that can be used to find RNA in genetic sequences
Problem with Rfam:

1) narrow families
2) pseudogenes
3) spliced RNA
4) speed
5) motif discovery

Task 3: faster search
Ravenna:. Can be somewhat slower than the BLAST/CM combo used in Rfam, but finds all the hits that the CM would on it's own and typically 100x faster.