Supplementary Notes for "RNA Motif Discovery" Lecture (12/3/2007)
Bart Trzynadlowski

- Pairs of mutations (compensatory) may reveal a motif
  - Algorithms for alignment may miss this
  - A double penalty will be assessed; poor alignment score results
- When evolutionary distance is close, amount of compensatory mutations is low
  - Algorithms like ClustalW work well (see Fig. 1)
  - As evolutionary distance increases (and therefore, compensatory mutations), alignment suffers, and as a result, algorithm accuracy declines

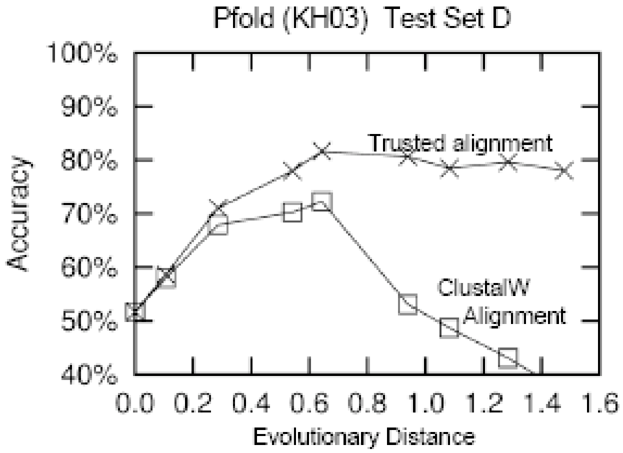

Fig. 1. Evolutionary distance vs. accuracy showing effect of poor alignment.

- CMFinder overview: Fig. 2.
  - Loop in the middle is just the EM algorithm
  - Loop constructs a Covariance Model, realigns, and then tries again
  - CMFinder has quite good accuracy on Rfam database families (Fig. 3)

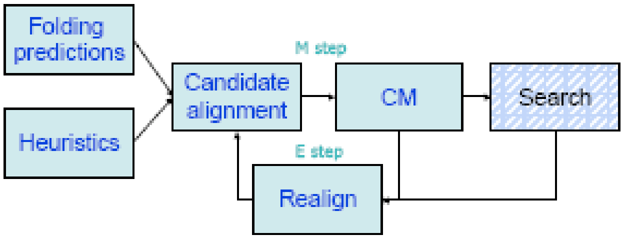

Fig. 2. Block diagram of CMFinder, from the lecture notes.

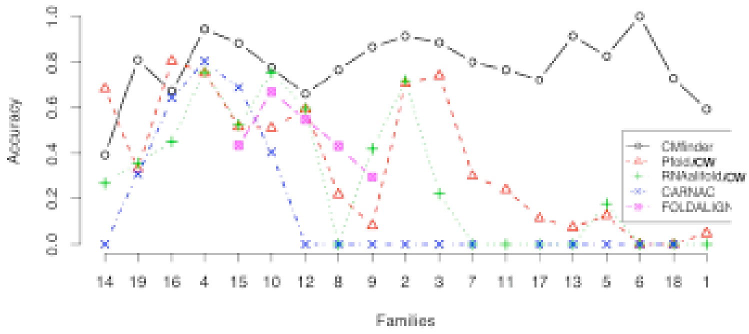

Fig. 3. CMFinder accuracy compared with other algorithms.

- Inferring parameters from alignments:
    - Pick structure that maximizes data likelihood
- Maximum likelihood structure, σ, maximizes $\sum_{(i,j)\in\beta} K_{ij}$, which is mutual information
    - Equal to $I_{ij} + \log\frac{p_{ij}}{s_i s_j}$
        - First term is mutual information term
        - Second (log) term is from folding calculation

- CMFinder cannot handle an entire genome, too slow
- CDD – Conserved Domain Database
    - “Domain” is some part of a protein that has a structure and performs a function
    - Use CDD to find similar proteins in different bacteria (find “upstream sequences”)
    - CMFinder will then spit out several motifs, take them & search for more

- Terminology alert: *cis-regulatory* means DNA near the gene it’s regulating.

- mRNA leader

    - Some bacteria use ~40% of their energy budget producing ribosomes. Therefore, proteins involved here should not be over- or under-produced (it would be wasteful and inefficient.) This is one possible mechanism.