- Pairs of mutations (compensatory) may reveal a motif
  o Algorithms for alignment may miss this
  o A double penalty will be assessed; poor alignment score results
- When evolutionary distance is close, amount of compensatory mutations is low
  o Algorithms like ClustalW work well (see Fig. 1)
  o 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.](image1)

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

![Fig. 2. Block diagram of CMFinder, from the lecture notes.](image2)
Fig. 3. CMFinder accuracy compared with other algorithms.

- Inferring parameters from alignments:
  o Pick structure that maximizes data likelihood

- Maximum likelihood structure, $\sigma$, maximizes $\sum_{(i,j) \in \beta} K_{ij}$, which is mutual information
  o 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
  o “Domain” is some part of a protein that has a structure and performs a function
  o Use CDD to find similar proteins in different bacteria (find “upstream sequences”)
  o 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
  o 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.