Selection of slides from:
Multiple Sequence Alignment

BMI/CS 576

www.biostat.wisc.edu/bmi576.html

Colin Dewey

cdewey@biostat.wisc.edu

Fall 2010

Multiple Sequence Alignment: Task Definition

- Given
  - a set of more than 2 sequences
  - a method for scoring an alignment
- Do:
  - determine the correspondences between the sequences such that the alignment score is maximized

Motivation for MSA

- establish input data for phylogenetic analyses
- determine evolutionary history of a set of sequences
  - At what point in history did certain mutations occur?
- discovering a common motif in a set of sequences (e.g. DNA sequences that bind the same protein)
- characterizing a set of sequences (e.g. a protein family)

Multiple Alignment of SH3 Domain

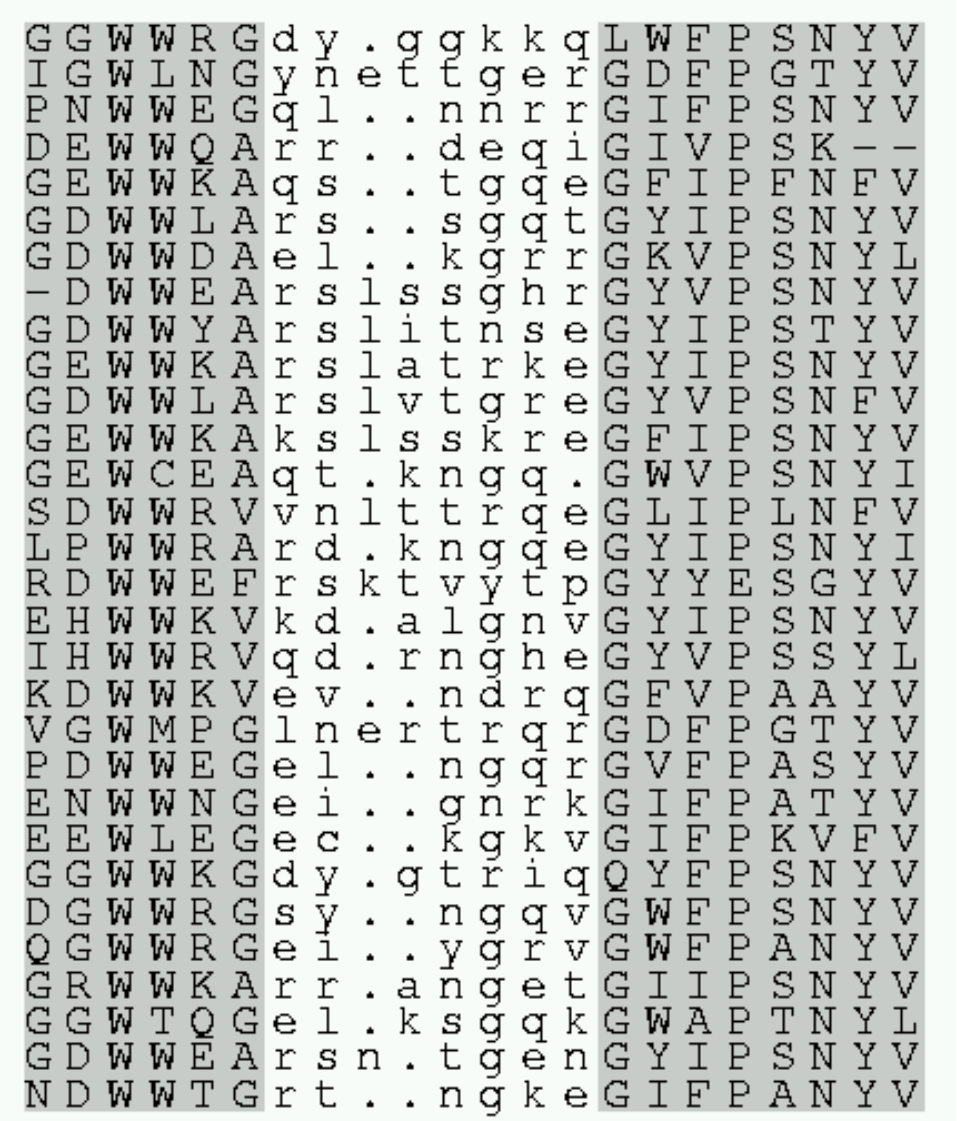

Figure from A. Krogh, An Introduction to Hidden Markov Models for Biological Sequences

Scoring a Multiple Alignment

- key issue: how do we assess the quality of a multiple sequence alignment?
- usually, the assumption is made that the individual *columns* of an alignment are independent

$$Score(m) = G + \sum_i S(m_i)$$

gap function

score of i^{th} column

- we'll discuss two methods
  - sum of pairs (SP)
  - minimum entropy

Scoring an Alignment: Sum of Pairs

- compute the sum of the pairwise scores

$$S(m_i) = \sum_{k<l} s(m_i^k, m_i^l)$$

$m_i^k =$ character of the *k*th sequence in the *i* th column

$s =$ substitution matrix

Scoring an Alignment: Minimum Entropy

- basic idea: try to minimize the *entropy* of each column
- another way of thinking about it: columns that can be communicated using few bits are good
- information theory tells us that an optimal code uses $-\log_2 p$ bits to encode a message of probability p

Scoring an Alignment: Minimum Entropy

- the messages in this case are the characters in a given column
- the entropy of a column is given by:

$$S(m_i) = -\sum_a c_{ia} \log_2 p_{ia}$$

$m_i =$ the i th column of an alignment m

$c_{ia} =$ count of character a in column i

$p_{ia} =$ probability of character a in column i

Dynamic Programming Approach

- can find optimal alignments using dynamic programming
- generalization of methods for pairwise alignment
  - consider *k*-dimension matrix for *k* sequences (instead of 2-dimensional matrix)
  - each matrix element represents alignment score for *k* subsequences (instead of 2 subsequences)
- given *k* sequences of length *n*
  - space complexity is

$$O(n^k)$$

Heuristic Alignment Methods

- since time complexity of DP approach is exponential in the number of sequences, heuristic methods are usually used
- *progressive alignment*: construct a succession of pairwise alignments
  - star approach
  - tree approaches, like CLUSTALW
  - etc.

- iterative refinement
  - given a multiple alignment (say from a progressive method)
    - remove a sequence, realign it to profile of other sequences
    - repeat until convergence

Multiple Alignment Case Study: The Cystic Fibrosis Gene

- cystic fibrosis (CF)
  - recessive genetic disease caused by a defect in a single-gene
  - causes the body to produce abnormally thick mucus that clogs the lungs and the pancreas

- the cystic fibrosis conductance regulator (CFTR) gene
  - gene and its role in CF identified in 1989 [Riordan et al., *Science*]
  - most common mutation is called ΔF508; a deletion of a phenylalanine (F) at position 508 in the CFTR protein
  - the CFTR protein controls the movement of salt and water into and out of cells; mutations in CFTR block this movement, causing mucus problem

So What Does CFTR Do?
A CFTR Multiple Alignment

Figure from Riordan et al, *Science* 245:1066-1073, 1989.

Multiple Alignment Case Study: the Cystic Fibrosis Gene

- two key features of the protein made apparent in multiple sequence alignment (and other analyses)
  - membrane-spanning domains
  - ATP-binding motifs
- these features indicated that CFTR is likely to be involved in transporting ions across the cell membrane

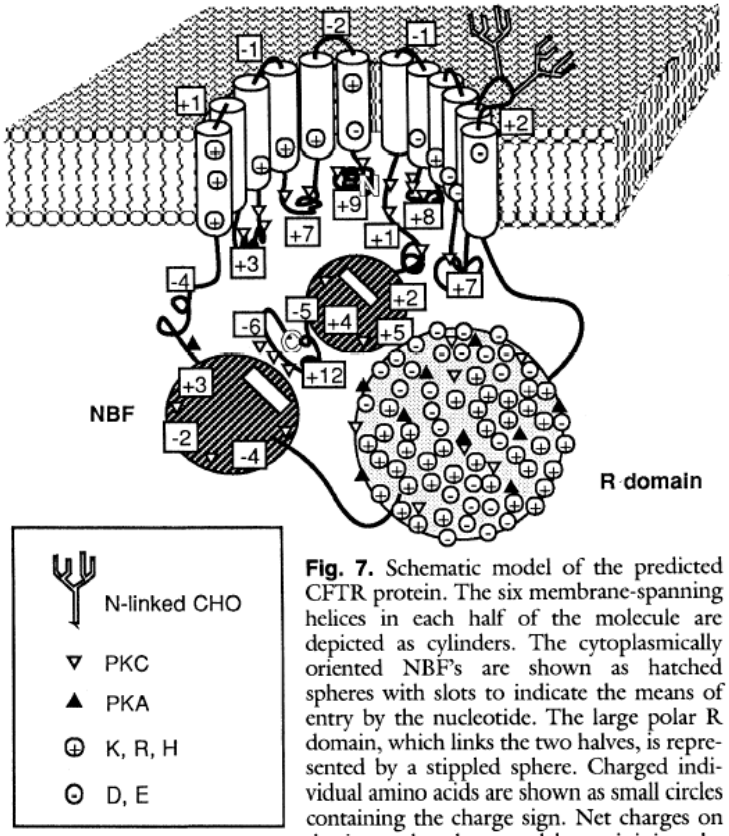

Fig. 7. Schematic model of the predicted CFTR protein. The six membrane-spanning helices in each half of the molecule are depicted as cylinders. The cytoplasmically oriented NBF's are shown as hatched spheres with slots to indicate the means of entry by the nucleotide. The large polar R domain, which links the two halves, is represented by a stippled sphere. Charged individual amino acids are shown as small circles containing the charge sign. Net charges on the internal and external loops joining the membrane cylinders and on regions of the NBF's are contained in open squares. Potential sites for phosphorylation by protein kinases A or C (PKA or PKC) and N-glycosylation (N-linked CHO) are as indicated. K, Lys; R, Arg; H, His; D, Asp; and E, Glu.

Figure from Riordan et al, *Science* 245:1066-1073, 1989.

Notes on Multiple Alignment

- as with pairwise alignment, can compute *local* and *global* multiple alignments
- dynamic programming is not feasible for most cases -- heuristic methods usually used instead