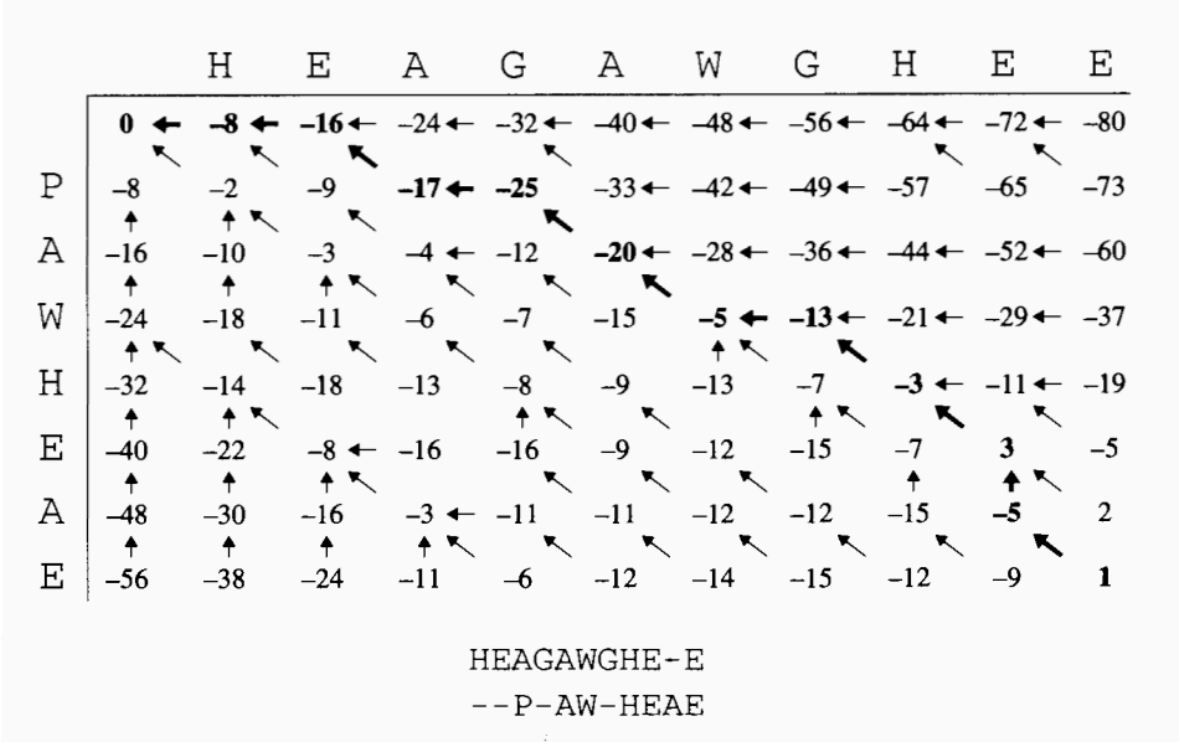
# Multiple Sequence Alignment

KEVIN LIN

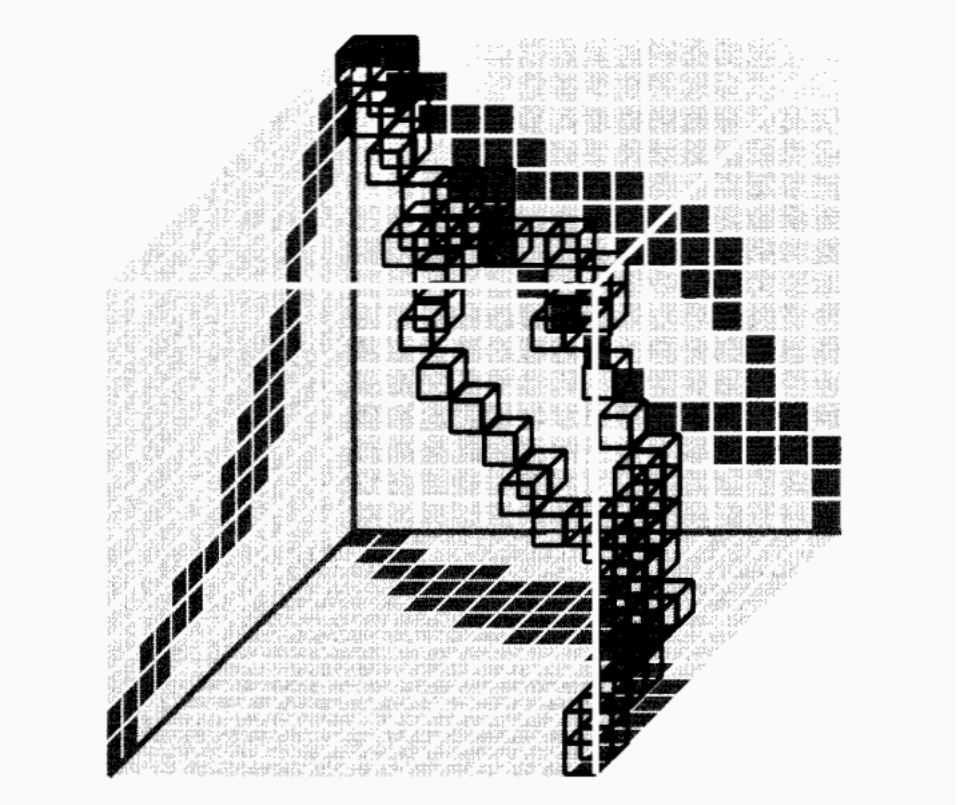
# Motivation

- Requirement for many other algorithms
- $F(i,j) = \begin{cases} F(i-1,j-1) + s_{x_i y_j} \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$
- Use of score matrix
- DP optimizes to $O(mn)$ time
- Extension to arbitrary dimensions $N$?



# Direct Extension

- $S(m_i)$ = score of column $i$, calculated by sums of pairwise scores
- $S(m) = \sum_i S(m_i)$, where $m$ is the total alignment score
- $$\alpha_{i_1,i_2,\ldots,i_n} = \begin{cases} \alpha_{i_{1-1},i_{2-1},\ldots,i_{n-1}} + S(x_{i_1}^1, x_{i_2}^2, \ldots, x_{i_N}^N) \\ \alpha_{i_1,i_{2-1},\ldots,i_{n-1}} + S(-, x_{i_2}^2, \ldots, x_{i_N}^N) \\ \alpha_{i_{1-1},i_2,i_{3-1}\ldots,i_{n-1}} + S(x_{i_1}^1, -, \ldots, x_{i_N}^N) \\ \ldots \\ \ldots \end{cases}$$
- $O(2^N L^N)$



# Progressive Approach - ClustalW

- Iteratively align 2 sequences at a time until 1 remain
- Decisions
  - Order of performing alignments
  - Linear or tree structure
  - Scoring sequences against alignments
- ClustalW
  - Compute distance matrix with pairwise alignment
  - Construct guide tree with neighbor join
  - Post-order alignment in order of decreasing similarity
  - Align alignments-alignments, alignments-sequences by average weighted sum of pairs

# Data

- Robin Gutell's Comparative RNA Website: RNA sequences
- Bioinformatics at Vrije Universiteit Brussel: amino acid sequences
- Simulate different sequence similarities by:
  - Setting seed sequence
  - Using HMM-like approach
  - States: follow seed sequence or not
  - Transition: based on sequence similarity
  - Epsilon error (insertion, deletion, mutation)

# Sources


- Durbin, Richard, et al. *Biological Sequence Analysis: Probabalistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 2013.
- http://www.rna.ccbb.utexas.edu/DAT/3C/Alignment/
- http://bioinformatics.vub.ac.be/databases/databases.html