CLUSTAL : a package for performing multiple sequence alignment on a microcomputer

Desmond G. Higgins and Paul M. Sharp

August 1988

Presented by : Rishika Agarwal

Outline:

- Introduction
- Previous Work
- Algorithm
- Results and Discussion

Outline:

- Introduction
- Previous Work
- Algorithm
- Results and Discussion

Introduction

- An approach for performing multiple alignments of large numbers of ammo acid or nucleotide sequences is described.
- The method is fast and memory-efficient to be easily implemented on a microcomputer.
- The results obtained are comparable to those from packages requiring mainframe computer facilities.

Outline:

- Introduction
- Previous Work
- Algorithm
- Results and Discussion

- For pairwise alignment, the dynamic programming algorithm of Needleman and Wunsch (1970) has been extensively used
  - guaranteed to maximise the overall similarity between the two sequences for any given gap penalty.
  - heavy computer time and space (core memory) requirements

- Approximate, heuristic approaches, such as that of Wilbur and Lipman (1984), have become popular, when a large number of pairwise comparisons are being made.
  - fast, and yield virtually the same results as the exact methods, as long as the sequences are not too dissimilar.

Outline:

- Introduction
- Previous Work
- Algorithm
- Results and Discussion

Algorithm

The algorithm consists of three stages:

1. Calculation of all pairwise sequence similarities

2. Construction of a dendrogram from the similarity matrix generated in stage 1

3. Multiple alignment of the sequences in a pairwise manner, following the order of clustering in the dendrogram from stage 2.

Step 1: **Calculation of all pairwise sequence similarities**

- Need to compute an NxN pairwise similarity matrix.
- N(N-1)/2 unique entries to be filled.
- Wilbur-Lapman algorithm used for fast computation of the matrix
- The scores are calculated as the number of exactly matching residues between two sequences in the optimal alignment, minus a penalty for every gap

Wilbur-Lapman algorithm for pairwise sequence similarity (1983)

- Gamma is an f-sequence

$$\text{score}(\Gamma) = \sum_{k=1}^{l} s(f^k) - \sum_{k=1}^{l-1} g(i_{k+1} - r_k - i_k - 1, j_{k+1} - r_k - j_k - 1) \tag{2.1}$$

where for each k, r_k is the length of f^k. The similarity, $S(A, B)$, is then defined as

$$S(A, B) = \max \{\text{score}(\Gamma) | \Gamma \text{ an } f\text{-sequence for } A \text{ and } B\}. \tag{2.2}$$

- Algorithm takes as input a set of alignment fragments F (cardinality of F = M).
- The time complexity mainly depends on M
- Several modifications of basic algorithm have been suggested for improving the time complexity
- Much faster than Needleman-Wunsch

Step 2: **Construction of a dendrogram**

- UPGMA is used fordengrogram construction, as it has good time and space requirements
- The dendrogram is of interest in its own right and can be used as input to a utility program for display on the screen
- The time required to construct a dendrogram for 100 sequences is approximately 3 min on a microcomputer.

Step 3: **Multiple Alignment**

- The sequences are taken, and aligned using the Wilbur and Lipman (1984) method, following the order of clustering in the dendrogram from step 2
- After each alignment, the aligned sequences with gaps inserted at appropriate positions (dashes) and a consensus sequence are written to a work file.
- The consensus sequence is used to represent the entire subtree or cluster of sequences under it

Multiple Alignment : **consensus sequence**

- First approach : Exact match
  - The consensus sequence contains only the residues found at a given position in all sequences, otherwise an unknown residue (‘x’) is recorded

- Second approach : Less sensitive to exact matches
  - conservative substitutions in protein alignments
  - a small degree of mismatch in the consensus sequences.

Conservative Substitutions in Protein Alignments

- A four-tier weighting scheme, based on the log-odds matrix of Dayhoff (1978) used to differentially weight aligned residues in protein alignments
- User-defined cut-off point for deciding whether or not a substitution is conservative.
- the four classes of match and their weights are:

(I) unmatched residues with a Dayhoff (1978) score of < 'cut-off' : score 0

(II) unmatched residues with a Dayhoff (1978) score > or = 'cut-off' : score 1

(III) exactly matched residues (except Cys, Phe, Trp or Tyr) : score 2

(IV) exactly matched Cys, Phe, Trp or Tyr : score 3

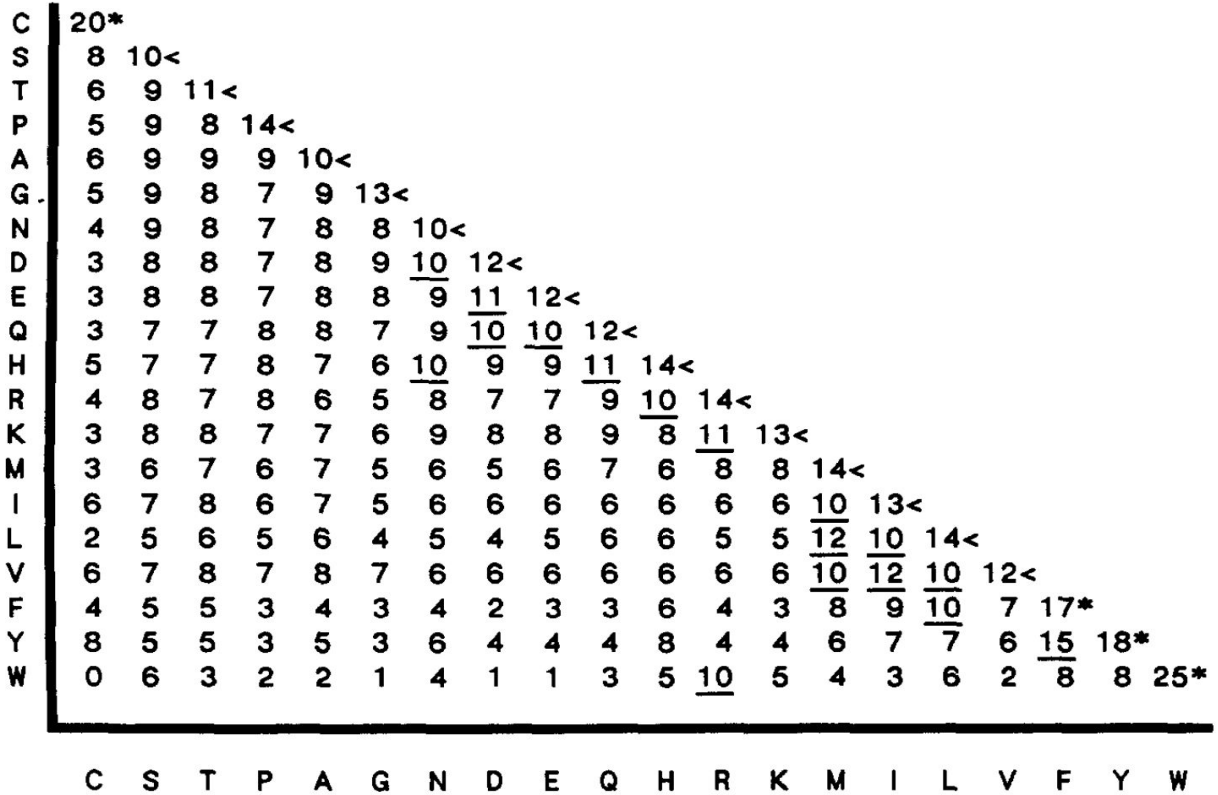

Fig. 2. Log-odds amino acid similarity matrix of Dayhoff (1978). The values in the matrix (scaled to lie between 0 and 25) indicate the frequency of substitution of any amino acid by another in related sequences. The single-letter amino acid code is used. The four-tier weighting system described in MATERIALS AND METHODS, section b*3i* (conservative substitutions), is illustrated. When a cut-off of 10 is used to decide whether or not a replacement is conservative, the four weights are indicated by: N (score = 0); $\underline{N}$ (score = 1); $N<$ (score = 2); N^* (score = 3), where N is a value from the matrix.

Partial Consensus sequences

- Include a residue in the consensus if it occurs in more than 75 % of the sequences that the consensus represents, otherwise it is recorded as an ‘unknown’ residue

Outline:

- Introduction
- Previous Work
- Algorithm
- Results and Discussion

Results

- Alignments
- Dendrograms
- Execution speed
- Comparison with other methods

Alignments

The alignments are judged on 2 datasets:

1) 7 divergent members of the globin family
2) 34 5s ribosomal RNA sequences

Alignment on Globins

- Lesk and Chothia (1980) identified seven α-helices, homologous between seven globin sequences
- The samples includes two mammalian α-globins, two β-globins, a myoglobin, a cyanohaemoglobin and a leghaemoglobin.
- Barton and Stemberg found that, although the sequences are highly diverged in terms of primary structure, they could correctly align all except two of the residues in each α-helix, across all the sequences.

Alignment on Globins

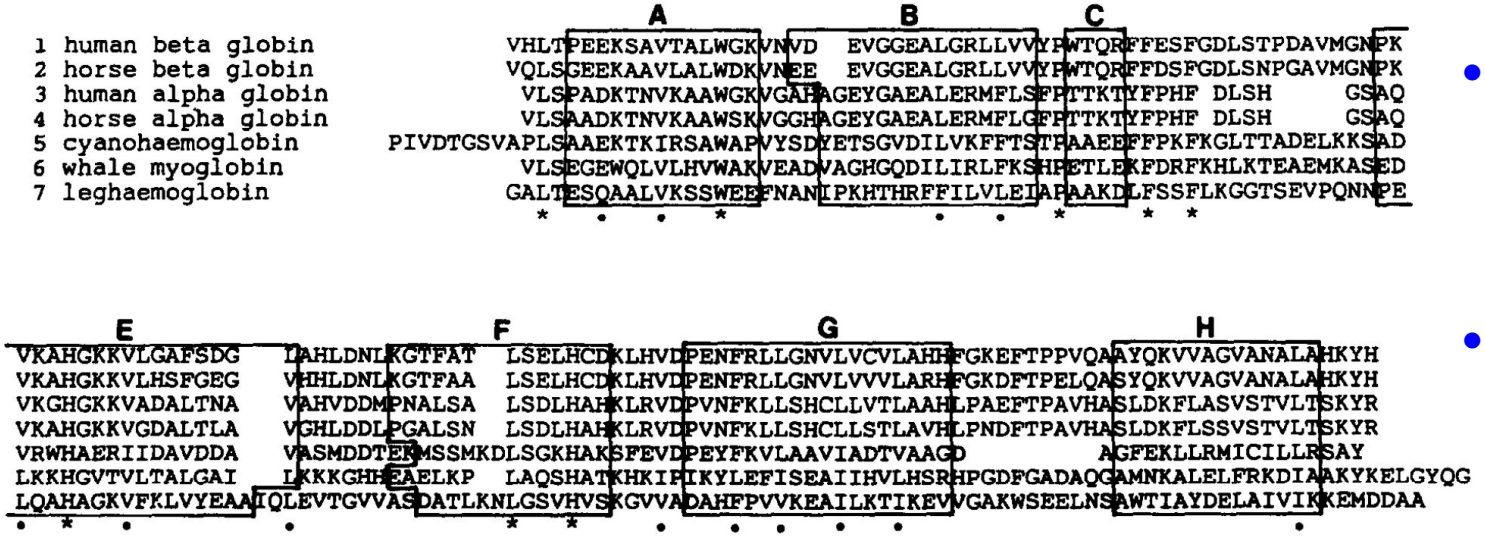

Fig. 3. CLUSTAL-produced multiple alignment of seven globin sequences taken from Lesk and Chothia (1980) (see RESULTS, section a*1*). Seven α-helices, homologous between all sequences, are labelled A to H and boxed. Asterisks indicate residues exactly conserved across all the sequences, while dots indicate regions where all pairs of residue have a similarity score greater than or equal to 10 (see Fig. 2).

- The α-helices are labelled A to H and, as can be seen, most of them are correctly aligned
- Only eight residues are **exactly** conserved in all of the molecules but the method still succeeds in aligning the regions of homologous secondary structure more or less correctly

- Barton and Stemberg (1987) found the same eight exactly conserved residues, but their method was slightly more accurate in aligning the α-helices

Alignment on 5S RNAs

- Hori et al. (1985) produced an alignment of 34 5s RNA sequences taken mainly from plants but including sequences from yeast, bacteria and a chloroplast.
- Their alignment was produced by **manually** aligning regions of known secondary structure (loops and base-paired regions in stems) followed by '**eyeball adjustment**'.
- Clustal succeeded in consistently aligning the regions of homologous secondary structure and produced an alignment not very different from that of Hori et al.

Results

- Alignments
- Dendrograms
- Execution speed
- Comparison with other methods

Dengrograms

- The dendrograms obtained from step 2 should not be used as phylogenetic trees:
    - Wilbur and Lipman alignment scores are only crude estimates of sequence similarity. No corrections are made for multiple substitutions or replacements.
    - UPGMA has been criticised as a method for inferring phylogenies from sequence data due to its inability to deal with unequal rates of evolution along different lineages. This may lead to errors in the tree topology

Dendrograms

- Dendrograms used should have biologically plausible topologies, since the order of clustering has an effect on the position of gaps in the final alignment.
- Dendrogram for **5S RNA sequences** shown
  - The sequences corresponding to the major plant groups (Bryophytes, Pteridophytes, Gymnosperms, etc.) cluster in an appropriate manner.
  - The topology is almost identical to that produced by Hori et al. using a more sophisticated strategy.
  - In general, CLUSTAL yields a satisfactory topology in a wide variety of situations.
  - In cases where this is not so, the user can manually adjust the topology or use an alternative approach.

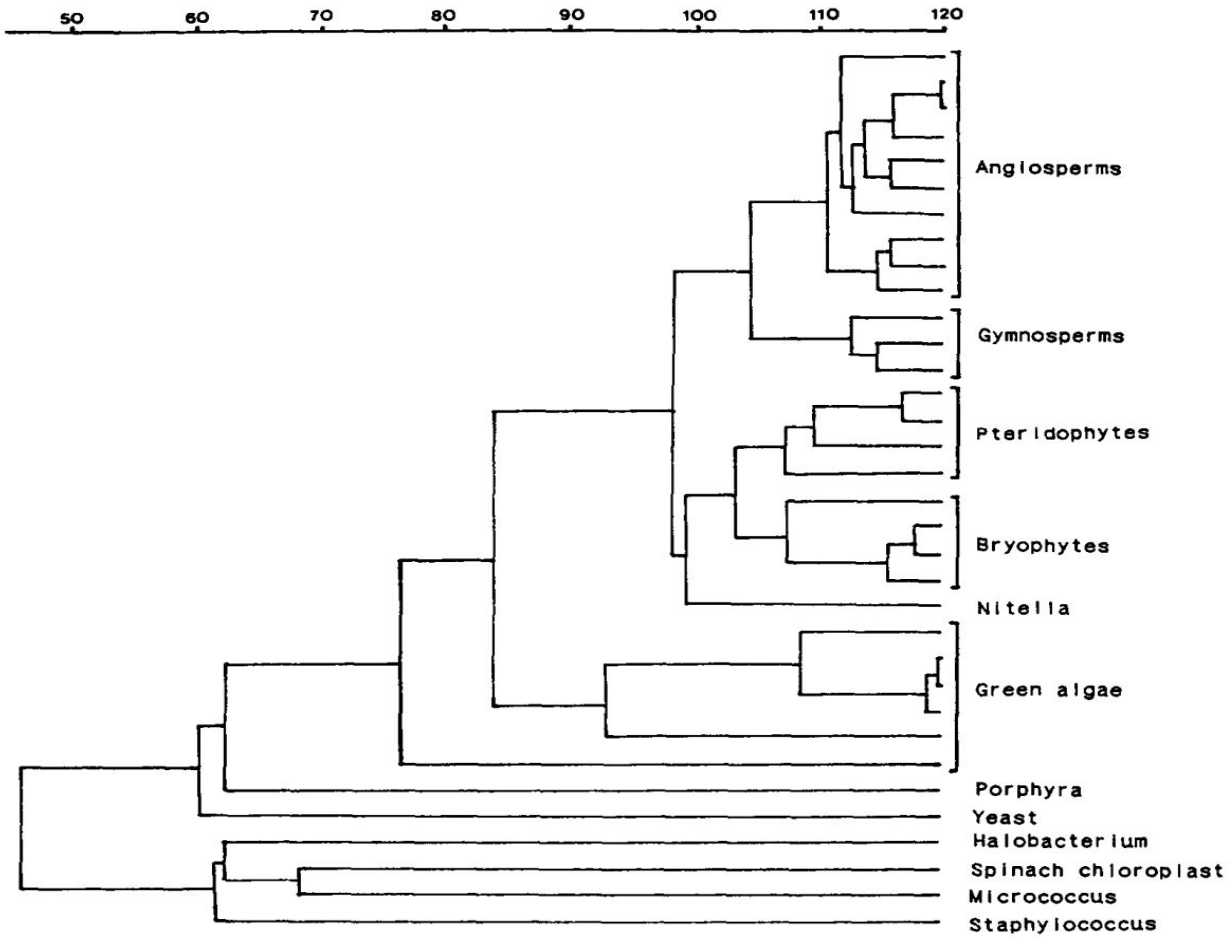

Fig. 4. UPGMA dendrogram of 34 plant, yeast and bacterial 5S RNA sequences. The sequences were taken from Hori et al. (1985) (see RESULTS, section a2). The major plant taxonomic groupings are indicated. The scale across the top margin shows the number of matching nucleotides (after alignment) between two clusters or sequences.

Results

- Alignments
- Dendrograms
- Execution speed
- Comparison with other methods

Execution Speed

- For large numbers of sequences, the greatest fraction of the time required for a multiple alignment is taken up in **calculating the pairwise sequence similarities** to construct the dendrogram.

- Both of the examples given previously (Globins, 5S RNAs) were carried out in less than 5 min on a microcomputer.

Results

- Alignments
- Dendrograms
- Execution speed
- Comparison with other methods

Comparison with other methods (of that time)

- For 2-sequence alignment : Needleman and Wunsch (1970)
- For 3-sequence : Murata et. al extended NW's pairwise alignment (1985)
- For 4 or more sequences : heuristics
  - Methods based on iteratively building a consensus sequence of all sequences to be aligned - Bains, 1986
  - Methods based on finding sub-sequences common in some or all sequences - Sobel and Martinez, Waterman (1986)
- Clustal inspired from Feng and Doolittle (1987) - series of pairwise alignments
- Taylor (1987) and Barton and Sternberg (1987) used ordered list of sequences derived from pairwise sequence distances to determine order of alignment

Notes on progressive pairwise alignment

- All pairwise methods are very fast and memory-efficient, so good at aligning large number of sequences.
- Placing gaps in a progressive manner reduces the complexity of the problem
- Feng and Doolittle justify the progressive alignment approach in terms of gap-placements.