Additional Paper for “Multiple Sequence Alignment”

**Paper reference:**

**Abstract:**
Since multiple sequence alignment is an NP-Hard problem, it is common practice to employ heuristic solutions to find satisfactory alignments. The current paper discusses an approach where a multiple alignment of k sequences is found which preserves k-1 pairwise alignments as specified by the edges of a given tree. It is shown that by using pairwise alignments which include consistency information from other sequences, very good accuracy can be obtained. The objective function is to find the shortest preserving alignment, which can be solved in polynomial time without using any heuristics.

**Discussion:**
The shortest preserving alignment approach aims to maximize the consistency between the resulting multiple alignment and a given set of pairwise alignments on aligned residue pairs. This approach tackles the problem when there are multiple alignments which preserve the given k−1 pairwise alignments. Moreover, once the tree and the pairwise alignments are fixed, no additional parameters or a scoring scheme for multiple alignments are needed.

The problem of multiple sequence alignment is reduced to finding a topological partial ordering on a directed acyclic graph where each vertex represents a partially aligned column and unordered vertices are allowed to share the same label. This problem can be solved in polynomial time by the following algorithm:

Given a tree T and a pairwise alignment $P_{ij}$ for each edge (i, j), the algorithm solves the shortest preserving multiple alignment problem in linear time by two successive graph reductions. The first reduction constructs an undirected graph $G = (V, E)$, where $E$ specifies exactly all the preservation constraints. The second reduction constructs a directed acyclic graph $G' = (V', E')$ from $G$ to specify ordering constraints. A multiple alignment can be obtained by finding a topological partial ordering in $G'$. Moreover, a topological partial ordering of $G'$ that uses the smallest number of labels specifies a shortest preserving multiple alignment. The overall time complexity of the entire procedure is $O(kn)$, where k is the number of pairwise alignments and each pairwise alignment is of length $O(n)$.

The algorithm has given comparable performance on three sets of benchmark multiple alignments: BAliBASE (1999), PREFAB (2004), and SABmark (2004). To further improve accuracy, various directions are proposed like improving the quality of the pairwise alignments and iteratively refining the solution after the preserving multiple sequence alignment is obtained.