2.2.4. Multiple Alignment of Protein Sequences

Multiple alignment is the basic methodology for analyzing sets of protein sequences, once they have been delineated using database searches, on the basis of a common function, or both. There is no general solution for the optimal multiple-alignment problem (unlike the pairwise alignment problem, for which the Smith-Warterman algorithm gives the exact solution), and therefore all the existing methods have to use heuristic approaches. These numerous methods may be divided into global and local, and on a different basis, into hierarchical and nonhierarchical approaches. Many of these methods have been recently reviewed and compared ***(38)***. Here, we present results produced by only one program, Multiple Alignment Construction and Analysis Workbench (MACAW) which is a local, nonhierarchical alignment method ***(39)***. In our experience, this method yields robust and useful alignments for a broad variety of sequence sets, some of which may be only very distantly related if a simple step-by-step procedure is followed. The MACAW algorithm initially detects all high-scoring, pairwise, local alignments in a set of sequences, and then expands them by adding similar segments from other sequences. The probability of each multiple alignment block to be found by chance is calculated using the Karlin-Altschul statistics. The program works in an interactive mode, allowing the user to select the regions of the sequences to be compared, as well as the alignment parameters and the minimal number of sequences in a block. With distantly related sequences, the best results are obtained if the block with the highest conservation is identified first and used as an anchor for subsequent detection of additional, compatible blocks, with weaker similarity in the remaining portions of the sequences. The decrease of the search space at each step, justified by the identification of the conserved block(s) at the preceding step, allows the detection of even very subtle signals. The end result produced by MACAW is a series of compatible multiple-alignment blocks separated by unaligned sequence segments.

Figure 6 schematically shows the alignment of the amino acid sequences of the capsid proteins of rod-shaped plant viruses produced by MACAW. This analysis confirms, at a statistically significant level, the existence of three conserved blocks detected previously ***(37)***, whereas the alignment in the regions separating them remains uncertain.

2.2.5. Protein Sequence Motifs

In many cases, screening of sequence databases for motifs conserved in protein families and superfamilies allows the detection of distant similarities that may not be detectable by methods for pairwise similarity search (e.g., BLAST). The simplest approach to motif analysis includes searching for simple, regular