proteins; similarity to known cellular proteins was not observed. Interpretation of the search results can be made on the basis of the pairwise score between the query sequence and a database sequence. Typically, the scores above 90 indicate genuine relationship. The so-called twilight zone of hits, with the score values ranging from 30 (current default-cutoff values in BLAST programs) to 70–90, may contain spurious hits, as well as potentially interesting similarities. **Figure 4A** shows a portion of the BLASTP output for the comparison of the TMV capsid protein sequence with the NR database. Only the alignments with the capsid proteins of other tobamoviruses stand out as highly statistically significant. Additional hits with virus capsid proteins, e.g., the *Nicotiana velutina* mosaic virus capsid protein sequence, and even potentially related nonstructural proteins from bymoviruses ***(6)***, may be detectable, but they are not statistically significant and are interspersed with completely irrelevant hits, like the one with the prolactin II precursor (**Fig. 4A**). Although the judgment of whether the twilight zone alignments are relevant can be based on the available biological information, rigorous evaluation of the significance of the hits requires methods for multiple alignment and motif analysis, as discussed below. Use of other methods for database search, namely FASTA and BLITZ, failed to reveal additional relationships between capsid proteins (data not shown).

When a group of functionally related proteins has diverse sequences, a useful trick that may allow one to detect subtle similarities is to limit the search space to the members of this particular group. **Figure 4B** shows the results of the comparison of the TMV capsid protein sequence with the representative set of plant virus capsid proteins sequences, i.e., with the database of 95 sequences mentioned above. Inspection of this output indicates that a conserved motif, previously described in the capsid proteins of rod-shaped viruses ***(37)***, can be detected using BLAST in several of these proteins, even though the use of this smaller search space still produced the results that were not statistically reliable.

The results of BLAST search can be used for classification of protein sequences. Apparently, the method of choice is single-linkage clustering that allows one to fully represent the relationships, even between distantly related sequences ***(34)***. According to this approach, a cluster is defined as a connected graph component, with each edge corresponding to a score above a chosen cutoff. It is required that each sequence in a cluster have a score above the cutoff, with at least one other sequence in the same cluster (and with no sequences outside the cluster), but not necessarily with all the members of the cluster. In order to cluster a set of sequences, it is first compared to itself using BLASTP (as implemented in the BLA program), and then the clusters are delineated on the basis of the BLAST scores (p values may be used alternatively). **Figure 5** depicts the results of clustering the plant virus capsid protein