neural network PHD program, in addition to secondary structure prediction, also predicts exposed and buried segments in proteins.

2.2. Sequence Similarity Analysis

2.2.1. Sequence Databases

The principal public databases of nucleotide sequences available for similarity search include GenBank, maintained at the National Center for Biotechnology Information (NCBI), NIH, Bethesda, MD ***(20)***; and EMBL from the European Bioinformatics Institute, Cambridge, UK ***(21)***. The two major protein databases are SWISS-PROT, at the University of Geneva, Switzerland ***(22)***; and PIR, at the National Biomedical Research Foundation, Washington DC ***(23)***. In addition, the NCBI supports the nucleic acid and protein version of the nonredundant (NR) database, which is produced by merging nonidentical entries from the above databases and is updated daily.

Special-purpose databases are typically created in the course of sequence-analysis projects. We constructed a database of plant virus capsid proteins by retrieving the entries containing the terms COAT and VIRUS, or CAPSID and VIRUS from the SWISS-PROT database. Several additional sequences of capsid proteins were extracted from the translated sequences in GenBank. Highly similar sequences, including in particular numerous potyvirus and potexvirus proteins, were removed from the database in order to retain representative sequences from each virus group. The resulting database of plant virus capsid proteins included 95 sequences. The sequences are in the simplest, FASTA format ***(24)***, which is accepted by most of the database-searching programs.

2.2.2. Database Searches

The most popular methods for database searches include FASTA ***(25,26)***, BLAST ***(27)***, and various implementations of the Smith-Waterman algorithm ***(28)***, such as BLITZ ***(29)***. The Smith-Waterman algorithm guarantees that the best local alignment, with gaps between the query sequence and each sequence in the database, is found, but encounters problems with statistical evaluation of alignments, even though solutions have been proposed recently ***(30)***. FASTA is based on an heuristic algorithm for finding gapped local alignments, following the initial finding of user-defined number of identical residues in a row.

BLAST (Basic Local Alignment Search Tool) finds high-scoring, ungapped, local alignments, and, for each of them, calculates the probability (p) that the alignment has been obtained by chance, using the extreme value distribution according to the Karlin-Altschul statistics ***(27,31,32)***. In addition, a heuristic approach has been developed to calculate p values for compatible multiple