Internet On-Ramp

The Sequence Manipulation Suite: JavaScript Programs for Analyzing and Formatting Protein and DNA Sequences

Paul Stothard, Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada (stothard@ualberta.ca)

JavaScript is an object-based scripting language that can be interpreted by most commonly used Web browsers, including Netscape® Navigator® and Internet Explorer®. In conjunction with HTML form elements, JavaScript can be used to make flexible and easy-to-use applications that can be accessed by anyone connected to the Internet (3). The Sequence Manipulation Suite (**http://www.ualberta.ca/~stothard/javascript/**) is a collection of freely available JavaScript applications for molecular biologists. It consists of over 30 utilities for analyzing and manipulating sequence data, including the following:

- **Codon Plot** accepts a DNA sequence and generates a graphical plot consisting of a horizontal bar for each codon. The length of the bar is proportional to the frequency of the codon in the codon frequency table you enter. Codon frequency tables for numerous organisms are available online (4) (**http://www.kazusa.or.jp/codon/**). Use Codon Plot to find portions of DNA sequence that may be poorly expressed, or to view a graphic representation of a codon usage table (by using a DNA sequence consisting of one of each codon type).
- **Codon Usage** accepts a DNA sequence and returns the number and frequency of each codon type. Since the program also compares the frequencies of codons that code for the same amino acid (synonymous codons), you can use it to assess whether a sequence shows a preference for particular synonymous codons.
- **CpG Islands** searches for CpG islands in a DNA sequence using the method of Gardiner-Garden and Frommer (2). CpG islands are often found in the 5′ regions of vertebrate genes, and therefore this program can be used to highlight potential genes in genomic sequences.
- **DNA/Protein Pattern Find** accepts a sequence along with a set of search patterns and returns the number and positions of sites that match the patterns. The search patterns can contain “wild cards”, which allow you to detect a variety of similar sequences using a single pattern.
- **DNA/Protein Stats** returns the number of occurrences of each residue in the sequence you enter. Percentage totals are also given for each residue, and for certain groups of residues, allowing you to quickly compare the results obtained for different sequences.
- **Filter DNA/Protein** removes non-protein or non-DNA characters from text. Use this program when you want to remove digits and blank spaces from a sequence to make it suitable for other applications. All programs in the Sequence Manipulation Suite automatically filter the sequences you enter before proceeding.
- **GenBank® Feature Extractor** accepts a GenBank file as input and reads the sequence feature information described in the feature table, according to the rules outlined in the GenBank release notes (**ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt**). The program concatenates the relevant sequence segments and returns each sequence feature in FASTA format. GenBank Feature Extractor is particularly helpful when you want to derive the sequence of a cDNA from a genomic sequence that contains many introns.
- **GenBank to FASTA** accepts a GenBank file as input and returns the entire DNA sequence in FASTA format. Use this program when you want to quickly remove all of the non-DNA sequence information from a GenBank file.
- **GenBank Trans Extractor** accepts a GenBank file as input and returns each of the protein translations described in the file in FASTA format. GenBank Trans Extractor should be used when you are more interested in the predicted protein translations of a DNA sequence than the DNA sequence itself.
- **Group DNA/Protein** adjusts the spacing of DNA or protein sequences and adds numbering. You can specify the group size (the number of residues per group) and the number of residues per line. The output of this program can serve as a convenient reference for use with some of the other programs in the Sequence Manipulation Suite (such as DNA/Protein Pattern Find) since the numbering and spacing allow you to quickly locate specific residues.
- **Ident and Sim** accepts a pair of aligned sequences (in FASTA format) and calculates their identity and similarity. Identity and similarity values are often used to assess whether or not two sequences share a common ancestor or function. Since many alignment programs do not calculate these values, you may find Ident and Sim quite useful when making pairwise comparisons.
- **Multiple Align Show** accepts a group of aligned sequences (in FASTA format) and formats the alignment to your specifications. You can specify the number of residues per line, and whether to use colored text or colored backgrounds to highlight matching residues. You can also set a consensus level, which specifies the percentage of residues that need to be identical in a column of the alignment for highlighting to be added. Use Multiple Align Show to enhance the output of sequence alignment programs.
- **ORF Finder** searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF along with its protein translation. Use ORF Finder to search newly sequenced DNA for potential protein-encoding segments.
- **Primer Show** accepts a DNA sequence along with a set of primer sequences and returns a textual map showing the annealing positions of the primers. The translation of the DNA sequence can be shown in the reading frames you specify. You can also choose the number of bases per line of the map, and whether to show the DNA in its single-stranded or double-stranded form. The primer sequences you enter can contain “wild cards”, a feature that allows Primer Show to handle degenerate primers. Use this program to produce a useful reference figure, particularly when you have designed a large number of primers for a particular template.

- **Protein Molecular Weight** accepts a protein sequence and calculates the molecular weight. You can append copies of commonly used epitopes and fusion proteins using the supplied list. Use Protein Molecular Weight when you want to predict the location of a protein of interest on a gel in relation to a set of protein standards.
- **Random DNA/Protein Sequence** generates a random sequence of the length you specify. Random sequences can be used to evaluate the significance of sequence analysis results.
- **Rest and Trans Map** accepts a DNA sequence and returns a textual map showing the positions of restriction endonuclease cut sites. The translation of the DNA sequence is also given, in the reading frame you specify. You can choose the number of bases per line of the map, and whether to show the DNA in its single-stranded or double-stranded form. Use the output of this program as a reference when planning cloning strategies, particularly when you need to consider the reading frames of the molecules you are ligating.
- **Restriction Summary** accepts a DNA sequence and returns the number and positions of restriction endonuclease cut sites. Use this program when you want to quickly determine whether or not an enzyme cuts a particular segment of DNA, and to produce a table to complement the output of Rest and Trans Map.
- **Reverse Complement** converts a DNA sequence into its reverse, complement or reverse-complement counterpart. You may want to work with the reverse complement of a sequence if it contains an ORF on the reverse strand.
- **Reverse Translate** accepts a protein sequence and uses a codon usage table to generate a graph that can be used to find regions of minimal degeneracy at the nucleotide level. Use Reverse Translate to design PCR primers for amplifying genes that encode similar proteins.
- **Shuffle DNA/Protein** randomly shuffles a sequence. Shuffled sequences can be used to evaluate the significance of sequence analysis results, particularly when sequence composition is an important consideration.
- **Simple Plot** calculates DNA sequence composition using a sliding window. The results are returned as a set of x and y values. The x-value is the position of the first base in the window of bases used in the calculation, and the y-value is the result of the calculation. Use Simple Plot when you are looking for segments of a DNA sequence with a particular base composition, rather than a specific sequence of bases.
- **TestCode** accepts a DNA sequence and calculates the TestCode value as described by Fickett (1). The reading frame of the DNA sequence does not need to be considered when performing this analysis. Use TestCode to predict whether or not a sequence encodes a protein.

In addition to the features described above, the interface to each application displays (where applicable) the genetic code, restriction enzyme set and codon usage table it uses so that you can make changes before performing an operation. Each program also comes with a complete set of default inputs to help illustrate the desired data formats.

To access a program in the Sequence Manipulation Suite, simply point your Web browser to (**http://www.ualberta.ca/~stothard/javascript/**) and click on the name of the program you want to use. The time needed for the applications to complete their tasks is dependent on how much input you give them. Try short sequences first to gauge how they perform on your computer.

Each program in the Sequence Manipulation Suite writes its results to the same output window, which appears when the first analysis is performed. Whenever you perform an additional analysis, the results are simply appended to the existing output data. By scrolling up and down in the window, you can easily compare the results of a particular operation performed on a series of sequences, or you can compare the results of many different operations performed on a single sequence. You can also copy a sequence segment of interest from the output window and paste it back into a Sequence Manipulation Suite program for further analysis or formatting.

To save or print the contents of the output window, copy the contents and paste them into your favorite text editor. If the spacing of the text looks incorrect after it has been pasted, switch to a fixed-width font such as Courier. The color and font of the text shown in the output window may not be conserved after copying and pasting, depending on which Web browser and text editor you are using.

The Sequence Manipulation Suite can be used online, or a copy of the programs can be downloaded and saved on your own computer for use offline (**http://www.ualberta.ca/~stothard/javascript/download.html**). One advantage of using a local copy of a program is that its default settings can be modified and saved to suit your preferences. For example, you might want to replace the codon usage tables with a table generated for the organism you study. You might also want to replace the restriction enzyme sets with a set consisting of the enzymes you have access to in your lab. Anyone with HTML experience is encouraged to make these changes. Readers familiar with JavaScript may also want to build new programs by combining and modifying existing portions of the Sequence Manipulation Suite.

Additional questions or comments regarding the Sequence Manipulation Suite can be sent by e-mail to the author at stothard@ualberta.ca.

REFERENCES

1. **Fickett, J.W.** 1982. Recognition of protein coding regions in DNA sequences. Nucleic Acids Res. *10*:5303-5318.
2. **Gardiner-Garden, M. and M. Frommer.** 1987. CpG islands in vertebrate genomes. J. Mol. Biol. *196*:261-282.
3. **Horton, R.M.** 1999. Biological sequence analysis using regular expressions. BioTechniques *27*:76-78.
4. **Nakamura, Y., T. Gojobori and T. Ikemura.** 2000. Codon usage tabulated from the international DNA sequence databases: status for the year 2000. Nucleic Acids Res. *28*:292.