Sequence Manipulation Suite:
Version 2

The Sequence Manipulation Suite is a collection of JavaScript programs for generating, formatting, and analyzing short DNA and protein sequences. It is commonly used by molecular biologists, for teaching, and for program and algorithm testing.

Sequence Manipulation Suite:
About

The Sequence Manipulation Suite is written in JavaScript 1.5, which is a lightweight, cross-platform, object-oriented scripting language. JavaScript is now standardized by the ECMA (European Computer Manufacturers Association). The first version of the ECMA standard is documented in the ECMA-262 specification. The ECMA-262 standard is also approved by the ISO (International Organization for Standards) as ISO-16262. JavaScript 1.5 is fully compatible with ECMA-262, Edition 3.

Sequences submitted to the Sequence Manipulation Suite do not leave your computer and are instead manipulated by your web browser, which executes the JavaScript. The Sequence Manipulation Suite was written by Paul Stothard (University of Alberta, Canada). Send questions and comments to stothard@ualberta.ca.

Here are short descriptions of the programs that comprise the Sequence Manipulation Suite:

Format Conversion:

Combine FASTA - converts multiple FASTA sequence records into a single sequence. Use Combine FASTA, for example, when you wish to determine the codon usage for a collection of sequences using a program that accepts a single sequence as input.

EMBL to FASTA - accepts one or more EMBL files as input and returns the DNA sequence from each in FASTA format. Use this program when you wish to quickly remove all of the non-DNA sequence information from an EMBL file.

EMBL Feature Extractor - accepts one or more EMBL files as input and reads the sequence feature information described in the feature tables. The program extracts or highlights the relevant sequence segments and returns each sequence feature in FASTA format. EMBL Feature Extractor is particularly helpful when you wish to derive the sequence of a cDNA from a genomic sequence that contains many introns.

EMBL Trans Extractor - accepts one or more EMBL files as input and returns each of the protein translations described in the files in FASTA format. EMBL Trans Extractor can be used when you are more interested in the predicted protein translations of a DNA sequence than the DNA sequence itself.

Filter DNA - removes non-DNA characters from text. Use this program when you wish to remove digits and blank spaces from a sequence to make it suitable for other applications.

Filter Protein - removes non-protein characters from text. Use this program when you wish to remove digits and blank spaces from a sequence to make it suitable for other applications.

GenBank to FASTA - accepts one or more GenBank files as input and returns the entire DNA sequence from each in FASTA format. Use this program when you wish to quickly remove all of the non-DNA sequence information from a GenBank file.

GenBank Feature Extractor - accepts one or more GenBank files as input and reads the sequence feature information described in the feature tables, according to the rules outlined in the GenBank release notes. The program extracts or highlights the relevant sequence segments and returns each sequence feature in FASTA format. GenBank Feature Extractor is particularly helpful when you wish to derive the sequence of a cDNA from a genomic sequence that contains many introns.

GenBank Trans Extractor - accepts one or more GenBank files as input and returns each of the protein translations described in the files 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.

One to Three - converts single letter translations to three letter translations.

Range Extractor DNA - accepts one or more DNA sequences along with a set of positions or ranges. The bases corresponding to the positions or ranges are returned, either as a single new sequence, a set of FASTA records, uppercase text, or lowercase text. Use Range Extractor DNA to obtain subsequences using position information.

Range Extractor Protein - accepts one or more protein sequences along with a set of positions or ranges. The residues corresponding to the positions or ranges are returned, either as a single new sequence, a set of FASTA

records, uppercase text, or lowercase text. Use Range Extractor Protein to obtain subsequences using position information.

Reverse Complement - converts a DNA sequence into its reverse, complement, or reverse-complement counterpart. The entire IUPAC DNA alphabet is supported, and the case of each input sequence character is maintained. You may want to work with the reverse-complement of a sequence if it contains an ORF on the reverse strand.

Split Codons - divides a coding sequence into three new sequences, each consisting of the bases from one of the three codon positions.

Split FASTA - divides FASTA sequence records into smaller FASTA sequences of the size you specify. An optional overlap value can be used to create sequences that overlap.

Three to One - converts three letter translations to single letter translations. Digits and blank spaces are removed automatically. Non-standard triplets are ignored.

Sequence Analysis:

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. 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 one or more DNA sequences 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 - reports potential CpG island regions using the method described by Gardiner-Garden and Frommer (1987). The calculation is performed using a 200 bp window moving across the sequence at 1 bp intervals. CpG islands are defined as sequence ranges where the Obs/Exp value is greater than 0.6 and the GC content is greater than 50%. The expected number of CpG dimers in a window is calculated as the number of 'C's in the window multiplied by the number of 'G's in the window, divided by the window length. CpG islands are often found in the 5' regions of vertebrate genes, therefore this program can be used to highlight potential genes in genomic sequences.

DNA Molecular Weight - accepts one or more DNA sequences and calculates molecular weight. Sequences can be treated as double-stranded or single-stranded, and as linear or circular. Use DNA Molecular Weight when calculating molecule copy number.

DNA Pattern Find - accepts one or more sequences along with a search pattern and returns the number and positions of sites that match the pattern. The search pattern is written as a JavaScript regular expression, which resembles the regular expressions written in other programming languages, such as Perl.

DNA 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.

Fuzzy Search DNA - accepts a DNA sequence along with a query sequence and returns sites that are identical or similar to the query. You can use this program, for example, to find sequences that can be easily mutated into a useful restriction site.

Fuzzy Search Protein - accepts a protein sequence along with a query sequence and returns sites that are identical or similar to the query.

Ident and Sim - accepts a group of aligned sequences (in FASTA or GDE format) and calculates the identity and similarity of each sequence pair. Identity and similarity values are often used to assess whether or not two sequences share a common ancestor or function.

Mutate for Digest - accepts a DNA sequence as input and searches for regions that can easily be mutated to create a restriction site of interest. The program also reports protein translations so that you can see which reading frames are altered by the proposed mutations. Use Mutate for Digest to find sequences that can be converted to a useful restriction site using PCR or site-directed mutagenesis.

Multi Rev Trans - accepts a protein alignment and uses a codon usage table to generate a degenerate DNA coding sequence. The program also returns a graph that can be used to find regions of minimal degeneracy at the nucleotide level. Use Multi Rev Trans when designing PCR primers to anneal to an unsequenced coding sequence from a related species.

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. ORF Finder supports the entire IUPAC alphabet and several genetic codes. Use ORF Finder to search newly sequenced DNA for potential protein encoding segments.

Pairwise Align Codons - accepts two coding sequences and determines the optimal global alignment. Use Pairwise Align Codons to look for conserved coding sequence regions.

Pairwise Align DNA - accepts two DNA sequences and determines the optimal global alignment. Use Pairwise Align DNA to look for conserved sequence regions.
Pairwise Align Protein - accepts two protein sequences and determines the optimal global alignment. Use Pairwise Align Protein to look for conserved sequence regions.
PCR Primer Stats - accepts a list of PCR primer sequences and returns a report describing the properties of each primer, including melting temperature, percent GC content, and PCR suitability. Use PCR Primer Stats to evaluate potential PCR primers.
PCR Products - accepts one or more DNA sequence templates and two primer sequences. The program searches for perfectly matching primer annealing sites that can generate a PCR product. Any resulting products are sorted by size, and they are given a title specifying their length, their position in the original sequence, and the primers that produced them. You can use linear or circular molecules as the template. Use PCR Products to determine the product sizes you can expect to see when you perform PCR in the lab.
Protein GRAVY - Protein GRAVY returns the GRAVY (grand average of hydropathy) value for the protein sequences you enter. The GRAVY value is calculated by adding the hydropathy value for each residue and dividing by the length of the sequence (Kyte and Doolittle; 1982).
Protein Isoelectric Point - calculates the theoretical pI (isoelectric point) for the protein sequence you enter. Use Protein Isoelectric Point when you want to know approximately where on a 2-D gel a particular protein will be found.
Protein Molecular Weight - accepts one or more protein sequences and calculates molecular weight. You can append copies of commonly used epitopes and fusion proteins using the supplied list. Use Protein Molecular Weight when you wish to predict the location of a protein of interest on a gel in relation to a set of protein standards.
Protein Pattern Find - accepts one or more sequences along with a search pattern and returns the number and positions of sites that match the pattern. The search pattern is written as a JavaScript regular expression, which resembles the regular expressions written in other programming languages, such as Perl.
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.
Restriction Digest - cleaves a DNA sequence in a virtual restriction digest, with one, two, or three restriction enzymes. The resulting fragments are sorted by size, and they are given a title specifying their length, their position in the original sequence, and the enzyme sites that produced them. You can digest linear or circular molecules, and even a mixture of molecules (by entering more than one sequence in FASTA format). Use Restriction Digest to determine the fragment sizes you will see when you perform a digest in the lab.
Restriction Summary - accepts a DNA sequence and returns the number and positions of commonly used restriction endonuclease cut sites. Use this program if you wish to quickly determine whether or not an enzyme cuts a particular segment of DNA.
Reverse Translate - accepts a protein sequence as input and uses a codon usage table to generate a DNA sequence representing the most likely non-degenerate coding sequence. A consensus sequence derived from all the possible codons for each amino acid is also returned. Use Reverse Translate when designing PCR primers to anneal to an unsequenced coding sequence from a related species.
Translate - accepts a DNA sequence and converts it into a protein in the reading frame you specify. Translate supports the entire IUPAC alphabet and several genetic codes.
Sequence Figures:
Color Align Conservation - accepts a group of aligned sequences (in FASTA or GDE format) and colors the alignment. The program examines each residue and compares it to the other residues in the same column. Residues that are identical among the sequences are given a black background, and those that are similar among the sequences are given a gray background. The remaining residues receive a white background. You can specify the percentage of residues that must be identical and similar for the coloring to be applied. Use Color Align Conservation to enhance the output of sequence alignment programs.
Color Align Properties - accepts a group of aligned sequences (in FASTA or GDE format) and colors the alignment. The program examines each residue and compares it to the other residues in the same column. Residues that are identical or similar among the sequences are given a colored background. The color is chosen according to the biochemical properties of the residue. You can specify the percentage of residues that must be identical and similar for the coloring to be applied. Use Color Align Properties to highlight protein regions with conserved biochemical properties.
Group DNA - adjusts the spacing of DNA sequences and adds numbering. You can specify the group size (the

number of bases per group), as well as the number of bases per line. The output of this program can serve as a convenient reference, since the numbering and spacing allows you to quickly locate specific bases.
Group Protein - adjusts the spacing of protein sequences and adds numbering. You can specify the group size (the number of residues per group), as well as the number of residues per line. The output of this program can serve as a convenient reference, since the numbering and spacing allows you to quickly locate specific residues.
Primer Map - accepts a DNA sequence and returns a textual map showing the annealing positions of PCR primers. Restriction endonuclease cut sites, and the protein translations of the DNA sequence can also be shown. Use this program to produce a useful reference figure, particularly when you have designed a large number of primers for a particular template. Primer Map supports the entire IUPAC alphabet and several genetic codes.
Restriction 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. Use the output of this program as a reference when planning cloning strategies. Restriction Map supports the entire IUPAC alphabet and several genetic codes.
Translation Map - accepts a DNA sequence and returns a textual map displaying protein translations. The reading frame of the translation can be specified (1, 2, 3, or all three), or you can choose to treat uppercase text as the reading frame. Translation Map supports the entire IUPAC alphabet and several genetic codes.
Random Sequences:
Mutate DNA - introduces base changes into a DNA sequence. You can select the number of mutations to introduce, and whether or not to preserve the first and last three bases in the sequence, to reflect selection acting to maintain start and stop codons. The position of each mutation is chosen randomly, and multiple mutations can occur at a single site. Mutated sequences can be used to evaluate the significance of sequence analysis results.
Mutate Protein - introduces residue changes into a protein sequence. You can select the number of mutations to introduce, and whether or not to preserve the first residue in the sequence, to reflect selection acting to maintain a start codon. The position of each mutation is chosen randomly, and multiple mutations can occur at a single site. Mutated sequences can be used to evaluate the significance of sequence analysis results.
Random Coding DNA - generates a random open reading frame beginning with a start codon and ending with a stop codon. You can choose the genetic code to use and the length of the sequence to generate. Random sequences can be used to evaluate the significance of sequence analysis results.
Random DNA Sequence - generates random sequences of the length you specify. Random sequences can be used to evaluate the significance of sequence analysis results.
Random DNA Regions - replaces regions of DNA sequences with random bases. Random sequences can be used to evaluate the significance of sequence analysis results.
Random Protein Sequence - generates random sequences of the length you specify. Random sequences can be used to evaluate the significance of sequence analysis results.
Random Protein Regions - replaces regions of protein sequences with random residues. Random sequences can be used to evaluate the significance of sequence analysis results.
Sample DNA - randomly selects bases from the guide sequence until a sequence of the length you specify is constructed. Each selected base is replaced so that it can be selected again.
Sample Protein - randomly selects bases from the guide sequence until a sequence of the length you specify is constructed. Each selected residue is replaced so that it can be selected again.
Shuffle DNA - randomly shuffles a DNA sequence. Shuffled sequences can be used to evaluate the significance of sequence analysis results, particularly when sequence composition is an important consideration.
Shuffle Protein - randomly shuffles a protein sequence. Shuffled sequences can be used to evaluate the significance of sequence analysis results, particularly when sequence composition is an important consideration.

Sequence Manipulation Suite:
Download

The Sequence Manipulation Suite. Copyright (C) 2000, 2004 Paul Stothard. This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version. This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details. You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston, MA 02111-1307, USA.
Download the Sequence Manipulation Suite (sms2.zip or sms2.tar.gz)

To mirror the Sequence Manipulation Suite: Extract sms2.zip or sms2.tar.gz and place the resulting sms2 directory into a directory from which your server will serve HTML files.
To run the Sequence Manipulation Suite locally: Extract sms2.zip or sms2.tar.gz and load index.html from the resulting sms2 directory into your web browser.

This version of the Sequence Manipulation Suite represents a complete re-write of the previous version. The new version is much faster and has many new features. The previous version of the Sequence Manipulation Suite can still be accessed.
Send questions and comments to stothard@ualberta.ca.

Reference:
http://www.protocol-online.org/tools/sms2/index.html Provided by (Protocol Online)