Pattern Matching and Replacement

Basic syntax:

(SEQUENCE-LIKE-PATTERN *pattern* IN *sequence*)
(SEQUENCES-LIKE-PATTERN *pattern* IN *sequence*)
(TEXT-LIKE-PATTERN *pattern* IN *text*)
(TEXTS-LIKE-PATTERN *pattern* IN *text*)

(REPLACE-PATTERN *pattern* IN *text* BY *replacement-text*)

where *sequence* is anything that can be interpreted as a sequence (e.g. a string, a gene, a protein, a set of same),[*] *text* is any string, and *pattern* is a string the nature of which is discussed below.

Simple patterns:
Any string of characters, excluding special characters (see below).

Example:

```
(SEQUENCES-LIKE-PATTERN "GGATCC" IN (SEQUENCE-OF A7120.chromosome))
```

Character sets and some special characters:

`.`	Any character
`\\d`	Any digit
`\\D`	Any non-digit
`\\w`	Any word character (letters and digits)
`\\W`	Any non-word character
`\\s`	Any space character (space, tab, and newline)
`\\S`	Any non-space character
[*abc*]	Set of characters
[^*abc*]	Set of excluded characters
[*a*-*z*]	Set of characters from first character to last

Examples:

```
(TEXT-LIKE-PATTERN "\\d\\d-\\w\\w\\w-\\d\\d\\d\\d"
    IN "LOCUS    ANGLNA    2225 bp    DNA     linear   BCT 12-SEP-1993")
```

Extracts the date from a locus line of a GenBank file

```
(TEXT-LIKE-PATTERN "[^ACGT]" IN (SEQUENCE-OF Cw?0002))
```

Returns positions of the sequence with nonstandard nucleotides

[*] PLEASE NOTE: At present, only simple strings work.

Repetition symbols

`?`	Previous element may be present or absent
`+`	Previous element may be present 1 or any number of times (choose maximum number of times)
`+?`	Previous element may be present 1 or any number of times (choose minimum number of times)
`*`	Previous element may be absent or present any number of times (choose maximum number of times)
`*?`	Previous element may be absent or present any number of times (choose minimum number of times)
`{`*n*`}`	Previous element must be present the *n* number of times
`{`*m*`,n}`	Previous element may be present anywhere from *m* to *n* number of times

Example:

```
(SEQUENCE-LIKE-PATTERN "C..C..C...C" IN (SEQUENCE-OF p-Ssr3184))
```

Finds amino acid sequence with spaced cysteines.

```
(SEQUENCE-LIKE-PATTERN " [acgt]*" IN
    " 1021 accacgaagt tgctactggt ggtcagtgcg agctaggctt ccgctttggt")
```

Other special symbols

`\\t`	tab
`\\n`	newline
`\\.`	Period (because . itself is special)
`\\+`	Plus (because + itself is special)
`\\*`	Asterisk (because * itself is special)
`^`	Element that follows must occur at beginning of string (not to be confused with ^ within a set designation, e.g. `[^ACGT]`)
`$`	Element that follows must occur at end of string
`( )`	Group (to be considered a single element in pattern matching)
`( )`	Remember these elements
`|`	Or

Problems

1. Write a statement that will find all recognition sites of the restriction enzyme *Sfi*I in the genome of *Anabaena* PCC 7120. The enzyme recognizes the sequence `GGCCNNNNNGGCC`, where `N` can be any nucleotide.
2. Write a statement that will find all consensus NtcA-binding sites in the genome of *Anabaena* PCC 7120. You can find a description of NtcA-binding sites in the notes for March 1.
3. Write a statement like the one above, but looking for NtcA-binding sites properly spaced from a downstream promoter.
4. Find all proteins that contain the four-amino acid pattern typical of DNA methyltransferases. "aPPb", where "a" represents serine, aspartate, or asparagines, and "b" represents tryptophan or tyrosine. List the descriptions of the proteins you found.
5. Improve the search of the previous problem by looking for only those amino acid sequences in which the four-amino acid motif is preceded by "xxn", where x represents a hydrophobic amino acid and n represents any amino acid. Is this search better than the last?
6. Find the context 4 nucleotides to either side of the DNA sequence CGATCG in the genome of *Synechocystis* PCC6803. Any pattern?
7. Find the length of the longest sequence in the genome of *Synechocystis* PCC6803 between two GCGATCGC sequences.
8. Look for an origin of replication in *Synechococcus* PCC 7942 as a region that has two sequences (TTTTCCACA) within 40 nucleotides of each other.
9. Determine whether there are any nonstandard nucleotides (i.e., not A, C, G, or T) within any gene of *Trichodesmium erythreum.*
10. Extract all the words from a BioBIKE error message (generate a good lengthy message if you have to). **All** the words, where a word is defined as something bound by a space, hyphen, parenthesis, etc.
11. Parse a GenBank file
    a. Go to GenBank and find the file for *glnA* from Nostoc PCC 7120. Download it as a text file and upload the file into your file space of BioBIKE.
    b. Read in the file as a list of lines:

       `(DEFINE` *name-of-list* `AS (FILE-TO-STRING-LIST` *file-name*`))`

    c. Write a loop that will print out each line of the file, preceded by a number. Follow this form:

```
(FOR-EACH line IN fill-in-something
     FOR n FROM 1
     (DISPLAY-LINE fill-in-something))
```

    d. Take a look at the first line of the list. Assign that first line to a variable.
    e. Write a statement that will extract from the first line the number of nucleotides (bp) in the file.
    f. Write a statement that will extract the date of submission of the sequence.
    g. Assign to a variable the contents of the first line that contains the word "promoter"
    h. Write a statement that will extract the coordinates of the given promoter.

i. Write a loop that will extract and save in a list the coordinates of all promoters identified in the documentation
j. Write a loop that will extract and save in a list the sequences of all promoters identified in the documentation.
k. Assign to a variable the first line after ORIGIN
l. Write a statement that will extract the sequence parts of the line and only those parts (no numbers or blanks) and save them as a list.
m. Write a statement that will take that list and JOIN them into a single sequence.
n. Write a statement that will convert the joined sequence to upper case (use STRING-UPCASE).
o. Find a way to consider the lines from 61 to the end and extract the sequence from it, yielding, in the end, a single upper-case sequence with no numbers or blanks.
p. Improve on the above code, by having it find the beginning of the sequence itself.

12. Find all candidate iron-sulfer proteins in *Anabaena* PCC 7120, taking advantage of the fact that such proteins should contain the motif of four cysteines (C), spaced 2, 2, and 3 amino acids apart.

13. Find a way to identify the largest open-reading frames encoded in the DNA of plasmid pNpD.