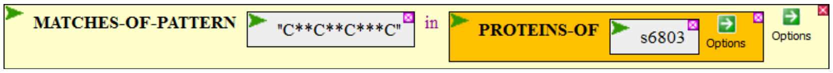
5/vi/13

BioBIKE Pattern Matching (using MATCHES-OF-PATTERN)

Character sets and some special characters:

`[abc]`	Set of characters
`[~abc]`	Set of excluded characters
`[a-z]`	Set of characters from first character to last
`[~a-z]`	Set of excluded characters from first character to last
`*`	Any character
#	Any digit (equivalent to [0-9])
~#	Any non-digit (equivalent to [~0-9])
$	Any word character (letters and digits) (equivalent to [0-9a-z])
~$	Any non-word character (equivalent to [~0-9a-z])
^ or ~@	Any space character (space, tab, and newline)
@ or ~^	Any non-space character
'	(Straight-quote) Either ' or "

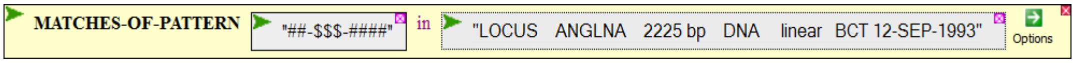
Examples:



MATCHES-OF-PATTERN "C**C**C***C" in PROTEINS-OF s6803

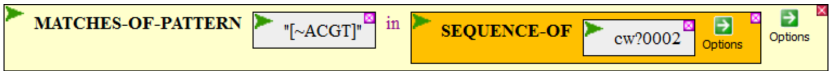
)

Looks for iron-sulfur cofactor binding site (four precisely placed cysteines) in sequences of the proteins of Synechocystis PCC 6803



MATCHES-OF-PATTERN "##-$$$-####" in "LOCUS ANGLNA 2225 bp DNA linear BCT 12-SEP-1993"

Looks within a locus line of a GenBank file for the date (two digits, hyphen, three letters, hyphen, four digits)



MATCHES-OF-PATTERN "[~ACGT]" in SEQUENCE-OF cw?0002

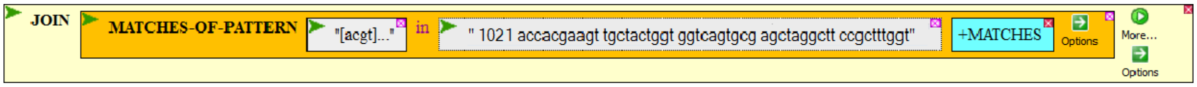
Looks within a gene sequence for nonstandard nucleotides (not A, C, G, or T)

Repetition symbols

{*n*}	Previous element must be present exactly *n* number of times
{*n*,}	Previous element must be present at least *n* number of times
{*m*,*n*}	Previous element may be present anywhere from *m* to *n* number of times
?	Previous element may be present or absent (equivalent to {0,1})
..	Previous element may be present 1 or any number of times (choose minimum number of times)
...	Previous element may be present 1 or any number of times (choose maximum number of times) (equivalent to {1,})

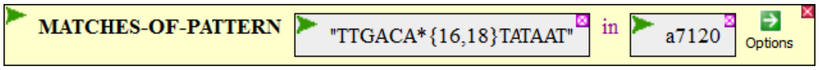
`?..`	Previous element may be absent or present any number of times (choose minimum length that satisfies the rest of the pattern)
`?...`	Previous element may be absent or present any number of times (choose maximum length that satisfies the rest of the pattern) (equivalent to {0,})

Examples:



```
JOIN MATCHES-OF-PATTERN "[acgt]..." in " 1021 accacgaagt tgctactggt ggtcagtgcg agctaggctt ccgctttggt" +MATCHES
```

Extracts contiguous blocks of nucleotides (a, c, g, or t) and joins them together, thereby ridding a GenBank sequence of numbers and spaces (there are easier ways!).



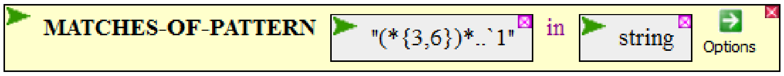
```
MATCHES-OF-PATTERN "TTGACA*{16,18}TATAAT" in a7120
```

Looks for consensus housekeeping promoter sequences in the genome of Anabgaena PCC 7120, defined as a perfect -35 sequence and perfect -10 sequence separated by 16 to 18 nucleotides.

Other special symbols

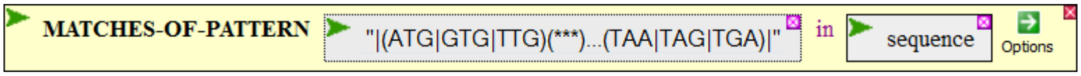
~	Negation
`	Back-quote/Escape (the character that follows is to be interpreted literally)
`#	Pound sign (because # itself is special)
`$	Dollar sign (because $ itself is special)
`*	Asterisk (because * itself is special)
`^	Carat (because ^ itself is special)
`*n*	Refers to a previously defined group, where *n* is the number of the group
( )	Group (to be considered a single element in pattern matching)
( )	Remember these elements
|	Bar (if at beginning of pattern, then indicates match starts at beginning of target) (if at end of pattern, then indicates match ends at end of target) (otherwise, indicates a choice between what precedes and what follows)

Example:



```
MATCHES-OF-PATTERN "(*{3,6})*..`1" in string
```

Looks for a segment 3 to 6 characters in length and a repetition of that segment after a gap of undetermined length. This pattern would therefore find matches in the strings "Walla Walla" or "Wallaby N. Wallace"



```
MATCHES-OF-PATTERN "|(ATG|GTG|TTG)(***)...(TAA|TAG|TGA)|" in sequence
```

Determines whether the sequence begins with a start codon (either ATG, GTG, or TTG), continues with any number of triplets, and ends with an in-frame stop codon (either TAA, TAG, or TGA).