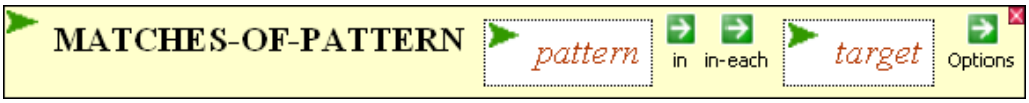
Pattern Matching

If you want to look for exact matches of a specific sequence (like all instances of "GAATTC" in a sequence) or almost exact matches (like the same but one mismatch is OK), then SEQUENCE-SIMILAR-TO will do the job. But sometimes your requirements are more complicated. Leaving bioinformatics for a moment, suppose you want to find all instances of social security numbers in a page of text. By eye it's easy – just scan for something of the form of ###-##-####, where # is a digit. You're looking not for a specific sequence of characters but rather a ***pattern***.

Similarly, it is often useful to find a pattern within a nucleotide or protein sequence. For example, certain proteins that catalyze oxidation-reduction reactions bind an iron-sulfer coenzyme through a series of cysteine amino acids that are spaced in a predicatable pattern `C...C...C..C`, where "." indicates the presence of some arbitrary amino acid. How do we look for such patterns?

Here's a way:



where *target* can be any string or sequence and *pattern* is a string the nature of which is discussed below.

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

Example:

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

Patterns with ambiguities:
Contains one or more of the characters below.

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:

```
(MATCHES-OF-PATTERN "C...C...C..C" IN candidate-gene)
```

Looks for iron-sulfer cofactor binding site in sequence of candidate gene

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

Looks for the date within a locus line of a GenBank file

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

Looks for nonstandard nucleotides within a gene sequence

It is possible to specify elements of a pattern of ambiguous length as well.

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:

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

Looks for blocks of nucleotides (any length), preceded by a space

```
(MATCHES-OF-PATTERN "am.{0,5}a \\w*" IN"I am not a crook (or am I?)")
```

Looks “am” within 5 characters of “ a “, followed by a sequence of letters of any length

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