FungiDB: Introduction to regular expressions

Regular expressions (RegEx) are sequences of characters used for pattern matching in text strings, e.g. 'aadgt', 'aa+dgt', 'a|d|c', '[mac]a'. Because nucleotide and amino acid sequences are text strings, regular expressions are very useful for finding motifs within sequences. There are differences in RegEx syntax between different tools or computer languages. For this section we will only discuss Perl RegExes.

Motifs often include repetitive or ambiguous assignments at some locations. The rules and special characters used in regular expressions help define the full set of strings that match the motif pattern.
The following is a description of some of these characters and examples of how they are used.
Although regular expressions seem complicated at first, they are very useful and easy to understand after going through some examples.

Special Characters

.	Match any character.
+	Matches "one or more of the preceding characters".
*	Matches "any number of occurrences of the preceding character", including 0.
?	Matches "zero or one occurrences of the preceding character".
[ ]	Matches any character contained in the brackets.
[^]	Match any character except those in the brackets.
{n}	Matches when the preceding character, or character range, occurs exactly n times.
{n,}	Matches when the preceding character occurs at least n times.
{n,m}	Matches when the preceding character occurs at least n times, but no more than m times.

The following codes can be used to represent classes of amino acids:

AA property	Amino acids	Code
Acidic	DE	0
Alcohol	ST	1
Aliphatic	ILV	2
Aromatic	FHWY	3
Basic	KRH	4
Charged	DEHKR	5
Hydrophobic	AVILMFYW	6
Hydrophilic	KRHDENQ	7
Polar	CDEHKNQRST	8
Small	ACDGNPSTV	9
Tiny	AGS	B
Turnlike	ACDEGHKNQRST	Z
Any	ACDEFGHIKLMNPQRSTVWY	.

Here are some examples of searches.

RegEx use case	Expected Result
ad+f (1 or more occurrences of 'd') would match any of the following	adf addf adddf addddddf
ad*f (0 or more occurrences of 'd') would match	af adf addf adddf
ad?f (0 or 1 occurrence of 'd') would match	af adf
a[yst]c would match	atc asc ayc
Specify residues.	R.L.[EQD] - an arginine (R), then any amino acid (.), then a leucine (L), then any amino acid (.), then either an aspartic acid, a glutamic acid, or a glutamine [EQD]
Specify the number of occurrences of a residue.	P{1,5} would match P from 1 to 5 times. .{1,30} would match any amino acid 1 to 30 times so you could find a motif within 30 amino acids of something like the beginning.
Pattern Anchors	^ Match only at the beginning of the string. $ Match only at the end of the string.
^mdef (e.g. a protein sequence starting with 'mdef') would match	mdef mdefab mdefaredfadfk but not match : edefa emdefa eeeemdef
kdel$ (searches for proteins ending with 'kdel', a standard ER retention signal) would match	eeeekdel kdel but not match : edefkdell akdeleefg