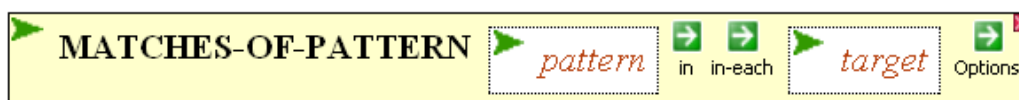


Pattern Matching

If you want to look for exact matches of a specific sequence (like all instances of “GAATTC” in a sequence) or almost exact matches (allowing some number of mismatches), then MATCHES-OF-ITEM or SEQUENCE-SIMILAR-TO (with MISMATCHES) will do the job. But sometimes your requirements are more complicated. Leaving bioinformatics for a moment, suppose you want to find all instances of dates in a page of text. You don't want stray numbers. 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*. While humans are exquisite pattern finders, computers can be taught to look for well specified patterns also.

SQ1. Try this. Go to the VCU main web page (www.vcu.edu). Select and copy the entire page. Then go to any BioBIKE instance and bring down MATCHES-OF-PATTERN:



Click the *target* entry box, and paste the page you copied into it, between a pair of quotation marks (it's easier if you type the quotation marks before pasting). Press Enter to close the entry box. Then, open the *pattern* entry box and enter a pattern for dates: "##/##/##" (not forgetting to press Enter). Execute the function. Did you find dates?

Why not? Hmm. It could be there are no dates of precisely that form. The first or second number might be one digit, not two. Maybe the year is given as four digits. Try again, using this pattern "#.../#.../#...", which allows for any number of digits at each position.

This problem may give you a taste for patterns and provide an example of a pattern that uses character sets (e.g. "#") and repetition symbols (e.g. "..."). To learn more about the capabilities of MATCHES-OF-PATTERN, see the [Help page](#) for the function and also the [list of special characters and sets](#).

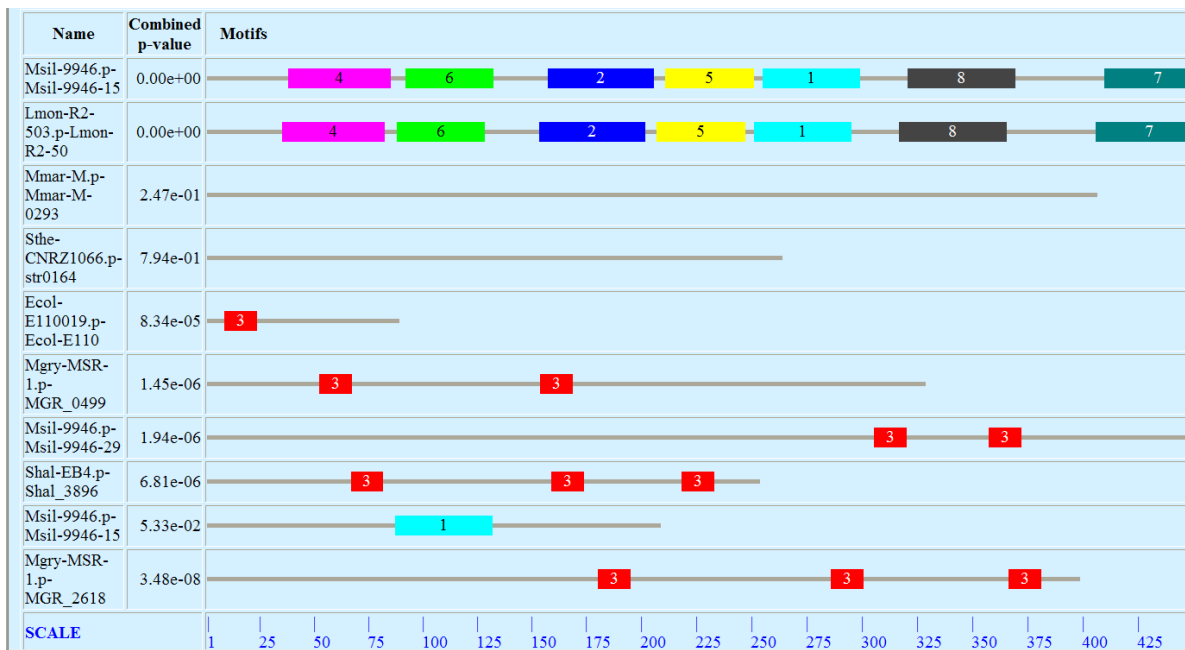
Similarly, it is often useful to find a pattern within a nucleotide or protein sequence. For example, suppose you're interested in certain proteins that catalyze oxidation-reduction reactions with the aid of an iron-sulfur coenzyme and want to be able to recognize their sequences. You could gather together a set of candidate proteins by looking for genes described by "iron-sulfur", but if you do, there's a good chance that in searching through the descriptions of every gene in every organism, you'll time out. Since you don't need an exhaustive list of such proteins but rather just a sampling, you can save time by getting the proteins from only a random subset of bacteria:



SQ2. Execute this function (in PhAnToMe/BioBIKE or replace *all-bacteria* with *all-cyanobacteria* and do it in CyanoBIKE). You'd like to get maybe 10-20 genes. Do you get them through this function?

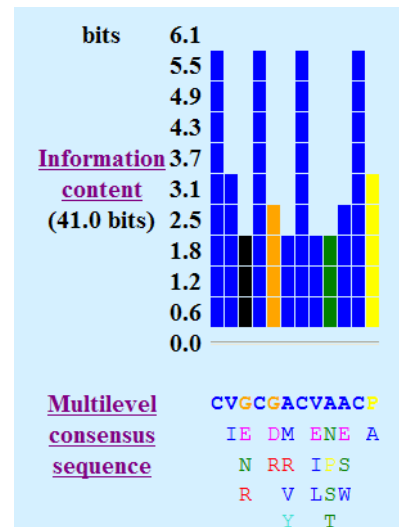
SQ3. Too many! Use the same CHOOSE-FROM trick to take a dozen of the genes at random, convert them to proteins using the PROTEINS-OF function, and find motifs within their sequences, using the MOTIFS-IN. Ask for 10 motifs, using the RETURN option. Once MOTIFS-IN returns the motifs page (usually 20-40 seconds), scroll down to the Summary of Motifs and look for the motif that occurs the most frequently.

Of course since you ran MOTIFS-IN with a quasi random set of iron-sulfur proteins, you'll get somewhat different results than I did, but you'll probably see something like this:



Scrolling up to the third motif, I find the remarkable consensus sequence and information profile shown at the right. In this motif, none of the amino acids are highly conserved except for four cysteines, and they're conserved 100%! It looks like these proteins annotated as iron-sulfur proteins have little similarity to one another except for a motif of C**C**C***C, where * denotes any amino acid.

Searching for this motif might lead me to other iron-sulfur proteins, but how to find them? Blast would be totally useless. It requires regions of contiguous similarity. Perhaps this is a job for MATCHES-OF-PATTERN? You think?



SQ4. Use MATCHES-OF-PATTERN to look for iron-sulfur oxidation-reduction proteins within the PROTEINS-OF a bacterium of your choosing or one taken at random (either in PhAnToMe/BioBIKE or CyanoBIKE). Use the pattern found above ("CC**C***C"). From the result, extract the proteins using the FIRST IN-EACH function, and display the DESCRIPTIONS-OF the proteins. Was the strategy effective in finding the desired proteins?**

So far, the patterns you've seen have made use of sets of characters (# = digits) and wild-cards (*), but pattern matching goes well beyond that. Perhaps now is the time to visit the [list of special characters and sets](#) used by pattern matching in BioBIKE. The first table shows the two character set characters you've already used (# and *), but you see that several others are defined and that it is possible to devise any character set you like, using the [] tool. For example, it may be useful to make up a character set [ACGT], the set of all possible nucleotides or its negative [~ACGT]. Here's an example of the utility of the latter:

SQ5. You've always had your doubts about the cyanobacterium *Synechococcus* BL107, which you've suspected of having a ratty genome. Test this out, using MATCHES-OF-PATTERN "[~ACGT]" to determine if there are any parts of the genome sequence composed of letters besides the four nucleotides. If you find such a region, follow it up by examining the sequence (using SEQUENCE-OF) to see if at least one of the position(s) identified by MATCHES-OF-PATTERN really does have a nonstandard letter.

The second table on the list shows ways of specifying repetition. For example:

SQ6. Look for candidate transcriptional terminators in phage TP901-1. You'll recall that a certain type of transcriptional terminator characteristically has a gapped palindrome followed by a run of at least 5 T's. The palindrome you'll need to recognize by eye (at least for now), but MATCHES-OF-PATTERN can certainly help. Search the set of downstream sequences of TP901-1 using a pattern consisting of 20 nucleotides followed by five or more T's ("{20}T{5,}"). Try this out with the SEQUENCE-DOWNSTREAM-OF one gene (say the first gene of TP901-1), then generalize to all genes of the phage. Find any reasonable candidates for transcriptional terminators?**