

# Protein Motif Searches and Regular Expressions

## 1. Using InterPro domain searches to identify unannotated kinesin motor proteins.

Note: For this exercise use <http://giardiadb.org>

Identify all genes annotated as hypothetical in all *Giardia* assemblages.

(Hint: use the full text search and look for genes with the word “hypothetical” in their product names)

### b. How many of these hypothetical genes have a kinesin-motor protein PFAM domain?

(Hint: add a step to the strategy.)

Go to the “Interpro Domain” search under similarity/pattern, start typing the work kinesin and it should autocomplete.)

**Identify Genes based on Text (product name, notes, etc.)**

Organism  select all | clear all | expand all | collapse all | reset to default  
 Giardia Assemblage A  
 Giardia Assemblage B  
 Giardia Assemblage E  
 select all | clear all | expand all | collapse all | reset to default

Text term (use \* as wildcard)

Fields  Alias  
 Cellular localization  
 Community annotation  
 EC descriptions  
 Gene ID  
 Gene notes  
 Gene product  
 GO terms and definitions  
 Protein domain names and descriptions  
 Similar proteins (BLAST hits v. NRDB/PDB)  
 User comments  
 select all | clear all

Advanced Parameters

**Add Step**

Run a new Search for  
 Transform by Orthology  
 Add contents of Basket  
 Add existing Strategy  
 Filter by assigned Weight  
 Transform to Pathways  
 Transform to Compounds

Genes  
 Genomic Segments  
 ORFs

Text, IDs, Organism  
 Genomic Position  
 Gene Attributes  
 Protein Attributes  
 Protein Features  
 Similarity/Pattern  
 Transcript Expression  
 Protein Expression  
 Cellular Location  
 Putative Function  
 Evolution  
 Population Biology

Protein Motif Pattern  
 InterPro Domain  
 BLAST

(Genes)

14987 Genes

Step 1

**Add Step 2 : InterPro Domain**

Organism  select all | clear all | expand all | collapse all | reset to default  
 Giardia Assemblage A  
 Giardia Assemblage B  
 Giardia Assemblage E  
 select all | clear all | expand all | collapse all | reset to default

Domain Database

Specific Domain(s)   
 PF06920 : Ded\_cyto Dedicator of cytokinesis  
 PF05804 : KAP Kinesin-associated protein (KAP)  
 PF00225 : Kinesin Kinesin motor domain

Free Text (use "" for wildcard)

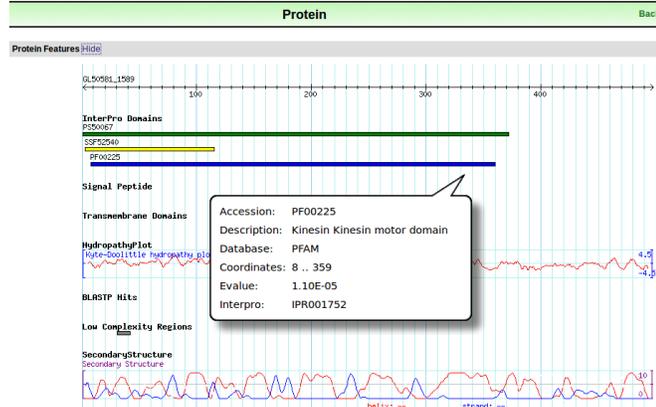
Advanced Parameters

**Combine Genes in Step 1 with Genes in Step 2:**

1 Intersect 2     1 Minus 2  
 1 Union 2     2 Minus 1  
 1 Relative to 2, using genomic colocation

- c. Go to the gene page for GL50581\_1589 and look at the protein feature section. Does this look like a possible motor protein?

Hint: click on the ID for GL50581\_1589 in the result table to go to the gene page. Scroll down to the protein section and mouse over the glyphs in the Protein Features graphic.



## 2. Using regular expressions to find motifs in TriTrypDB: finding active trans-sialidases in *T. cruzi*.

Note: for this exercise use <http://tritrypdb.org>

*T. cruzi* has an expanded family of trans-sialidases. In fact, if you run a text search for any gene with the word “trans-sialidase”, you return over 4000 genes among the strains in the database!!! Try this and see what you get.

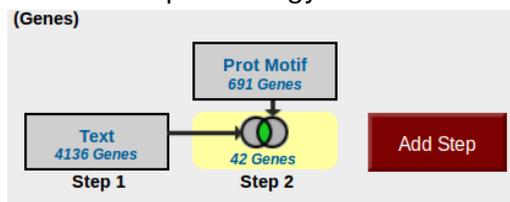
- b. However, not all of these are predicted to be active. It is known that active trans-sialidases have a signature tyrosine (Y) at position 342 in their amino acid sequence. Add a motif search step to the text search in ‘a’ to identify only the active trans-sialidases.

### Add Step 2 : Protein Motif Pattern

Hint: for your regular expression, remember that

you want the first amino acid to be a methionine, followed by 340 of any amino acid, followed by a tyrosine ‘Y’. Refer to [regular expression tutorial](#) if you need to.

If you need help, you can go to this sample strategy below to see the answer:



<http://tritrypdb.org/tritrypdb/im.do?s=a905e36f634f7b42>

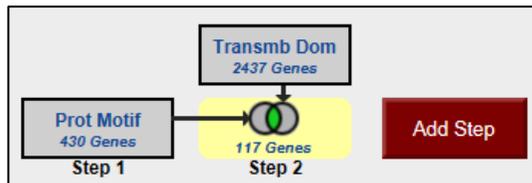
3. Using regular expressions to find motifs in CryptoDB: finding genes with the YXXΦ receptor signal motif

Note: for this exercise use <http://cryptodb.org>

The YXXΦ (Y=tyrosine, X=any amino acid, Φ=bulky hydrophobic [phenylalanine, tyrosine, threonine]) motif is conserved in many eukaryotic membrane proteins that are recognized by adaptor proteins for sorting in the endosomal/lysosomal pathway. This motif is typically located in the c-terminal end of the protein.

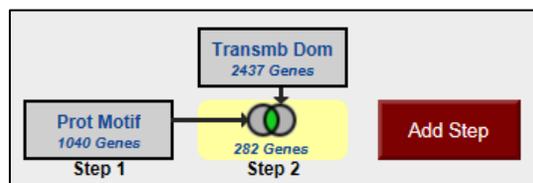
- a. Use the “protein motif pattern” search to find all *Cryptosporidium* proteins that contain this motif anywhere in the terminal 10 amino acids of proteins. (hint: for your regular expression, remember that you want the first amino acid to be a tyrosine, followed any two amino acids, followed by any bulky hydrophobic amino acid (phenylalanine, tyrosine, threonine). Refer to [regular expression tutorial](#) if you need to).

- b. How many of these proteins also contain at least one transmembrane domain.



- c. What would happen if you revise the first step (the motif pattern step) to include genes with the sorting motif in the C-terminal 20 amino acids? (hint: edit the first step and modify your regular expression)

Here is a saved strategy that provides you with the results of the above search:

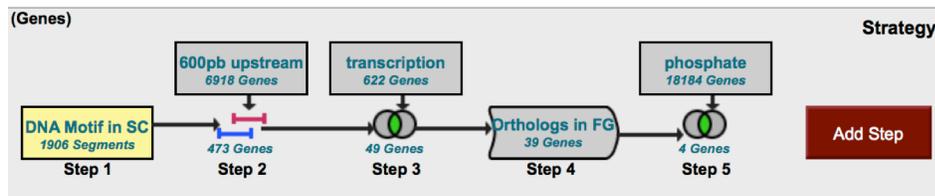


<http://cryptodb.org/cryptodb/im.do?s=928309b4c1b9ef3f>

## Exercise: Identify genes downstream of regulatory DNA motifs.

We will start our search with *S. cerevisiae* as a model organism where motifs and regulatory factors have been studied in great detail.

Transcriptional start sites are often located within a certain distance upstream of genes or gene clusters that they regulate. In some fungi, DNA motifs are important for regulation of processes linked to host cell invasion. Readily available genomic data facilitate the discovery of regulatory motifs and allow identification of novel genes via examination of orthologous sequences.



### 1. Initiate a search for CACGTG DNA motif

Click on “Search for genomic segments (DNA motif)” in the search menu at the top of the page.

Enter CACGTG in the search box and perform search in *Saccharomyces cerevisiae*.

Your search should return 1906 segments containing CACGTG motif. Next, we will look for putative regulatory targets of this motif by searching for genes located 600bp downstream of this sequence.

### 2. Identify genes with CACGTG motif being 600bp upstream of an open reading frame.

Click “Add Step”.

Choose Run a new search for Gene > Text > Organism and select “Relative to genomic location”. In the next screen you will set parameters of your search.

The screenshot shows the "Genomic Colocation" search interface. The search query is: "Return each Gene from Step 2 whose upstream region overlaps the exact region of a Genomic Segment in Step 1 and is on either strand". The "upstream region" is set to "Upstream: 600 bp". The "exact region" is set to "Exact".

Set up your search using the following guideline:  
*Return each gene from step 2 whose upstream region (600bp) overlaps the exact region of a Genomic Segment in Step1 (CACGTG) and is on either strand.*

Let's look a little closer at the 6918 genes we have identified and determine if there is an enrichment for certain biological processes.

Click on Analyze Results –BETA tab and view Gene Ontology based on biological processes.

Gene ID	Organism	Genomic Location	Product Description	Match Count	Region	Matched Regions
YAL068W-A	<i>S. cerevisiae</i> S288c	ScerS288c_Chr1: 538 - 792 (+)	Dubious open reading frame unlikely to encode a protein; identified by gene-trapping, microarray-bas...	2	-62 - 537 (+)	ScerS288c_Chr1:353-359f: 353 - 359 (+); ScerS288c_Chr1:353-359r: 353 - 359 (-)
YAL054C	<i>S. cerevisiae</i> S288c	ScerS288c_Chr1: 42,881 - 45,022 (-)	Acetyl-coA synthetase isoform which, along with Acs2p, is the nuclear source of acetyl-coA for histo...	4	45023 - 45622 (-)	ScerS288c_Chr1:45480-45486r: 45,480 - 45,486 (-); ScerS288c_Chr1:45402-45408r: 45,402 - 45,408 (-); ScerS288c_Chr1:45402-45408f: 45,402 - 45,408 (+); ScerS288c_Chr1:45480-45486f: 45,480 - 45,486 (+)
YAL053W	<i>S. cerevisiae</i> S288c	ScerS288c_Chr1: 45,899 - 48,250 (+)	Putative FAD transporter; required for uptake of FAD into endoplasmic reticulum; involved in cell wa...	4	45299 - 45898 (+)	ScerS288c_Chr1:45402-45408f: 45,402 - 45,408 (+); ScerS288c_Chr1:45480-45486r: 45,480 - 45,486 (-); ScerS288c_Chr1:45480-45486f: 45,480 - 45,486 (+); ScerS288c_Chr1:45402-45408r: 45,402 - 45,408 (-)
YAL038W	<i>S. cerevisiae</i> S288c	ScerS288c_Chr1: 71,786 - 73,288 (+)	Pyruvate kinase; functions as a homotetramer in glycolysis to convert phosphoenolpyruvate to pyruvat...	2	71186 - 71785 (+)	ScerS288c_Chr1:71264-71270f: 71,264 - 71,270 (+); ScerS288c_Chr1:71264-71270r: 71,264 - 71,270 (-)
YAL026C	<i>S. cerevisiae</i> S288c	ScerS288c_Chr1: 95,630 - 99,697 (-)	Aminophospholipid translocase (flippase) that maintains membrane lipid asymmetry in post-Golgi secre...	4	99698 - 100297 (-)	ScerS288c_Chr1:100152-100158r: 100,152 - 100,158 (-); ScerS288c_Chr1:100152-100158f: 100,152 - 100,158 (+); ScerS288c_Chr1:99877-99883r: 99,877 - 99,883 (-); ScerS288c_Chr1:99877-99883f: 99,877 - 99,883 (+)

### Gene Ontology Enrichment (2)

Find Gene Ontology terms that are enriched in your gene result. [Read More](#)

**Parameters**

Organism Fusarium graminearum PH-1

Ontology 
 Cellular Component  
 Molecular Function  
 Biological Process

GO Association Sources 
 Select all |  Clear all  
 InterPro predictions

P-Value Cutoff (0 - 1.0) 0.05

You can explore gene ontology enrichment for cellular components and molecular function.

It looks like this motif is located upstream of genes that belong to various biological processes. Let's narrow down our selection and look at those genes that are important for transcription.

Click “Add Step” and select Gene ID > Text and search for “transcription”.

Your search should return 39 genes that are assigned to 35 biological pathways.

How do you find which biological categories are enriched?

Hint: refer to Analyze Results tab > Gene Ontology > Biological Process

The screenshot shows a search interface with the following elements:

- Organism:** A tree view where **Saccharomyces cerevisiae S288c** is selected under the **Saccharomycetes** group.
- Text term (use \* as wildcard):** A search box containing the word "transcription".
- Fields:** A list of search fields with checkboxes, including **Alias**, **EC descriptions**, **Gene ID**, **Gene notes**, **Gene product**, **GO terms and definitions**, **Metabolic pathway names and descriptions**, **Phenotype**, **Protein domain names and descriptions**, **PubMed**, **Similar proteins (BLAST hits v. NRDB/PDB)**, and **User comments**.
- Parameters:** A section titled "Genes in Step 2 with Genes in Step 3:" with three Venn diagram options: **2 Intersect 3** (selected), **2 Union 3**, and **3 Minus 2**.

This search strategy can be extended to identify orthologous genes in less studied model systems. For example, let’s identify orthologs of the 49 *S. cerevisiae* genes in *Fusarium graminearum*, a plant pathogen.

Initiate search for orthologs in *F. graminearum* using “Add Step” tool. To determine which cellular components are enriched within our search click on “Analyze Results-BETA” tab and choose GO > Cellular Component

It is known that cellular components are affected by various growth conditions. For example, phosphate deprivation leads to production of phosphorus-free membranes in *F. graminearum*, which is an important process during host invasion.

In order to identify genes that have CACGTG DNA motif upstream of their reading frame and may participate in metabolism of phosphate, perform a text search for “phosphate”.

You should identify 4 genes that comprise a conserved PHO system in filamentous fungi, and have important roles during metabolic and cellular response to phosphate starvation.

