

## Sequence Exercises: Motifs, Domains and Colocation

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

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

- a. Identify all genes annotated as hypothetical in all *Giardia* assemblages. Use the full text search and look for genes with the word “hypothetical” in their product names.

The screenshot shows the InterPro search interface. The search criteria are: Organism: 6 selected, out of 6 (Giardia, Spironucleus); Text term: hypothetical; Fields: Gene product (checked). The search results show 18590 genes. A red arrow points from the 'Get Answer' button to the 'Add Step' button.

- b. How many of these hypothetical genes have a kinesin-motor protein PFAM domain?
- Add a step to the strategy. Go to the “Interpro Domain” search under ‘Protein features and properties’ similarity/pattern, start typing the work kinesin and it should autocomplete.

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

▼ Proteins Properties and Features [Download](#) [Data sets](#)

Transcript ID	Isoelectric Point	Molecular Weight	Has SignalP	Has TMHMM	Protein Length	Protein Browser
GL50581_4			no	no	320	<a href="#">Interactive</a>

Track details

Accession: PF00225

Description: Kinesin Kinesin motor domain

Database: PFAM

Coordinates: 45..113

Evalue: 2.50E-07

Interpro: IPR001752

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 9000 genes among the strains in the database!!! Try this and see what you get.
- Not all of the genes returned in (a) 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.

- Write a regular expression that defines a protein sequence that starts with a methionine, and is followed by 340 of any amino acids, followed by a tyrosine ‘Y’. Refer to [regular expression tutorial](#) if you need to.
- <http://tritrypdb.org/tritrypdb/im.do?s=0d7be75a64dbc2bb>

The image shows a workflow in the TriTrypDB interface. On the left, a box labeled "Text 9445 Genes Step 1" has a red "Add Step" button. A red arrow points from this button to a larger window titled "Add Step 2 : Protein Motif Pattern".

Inside the "Add Step 2" window, the "Pattern" field contains the regular expression `^M.{340}Y`. The "Organism" section shows a tree view where "Trypanosoma" is expanded and "Trypanosoma cruzi" is selected. Below this, there are options to "Combine Genes in Step 1 with Genes in Step 2":

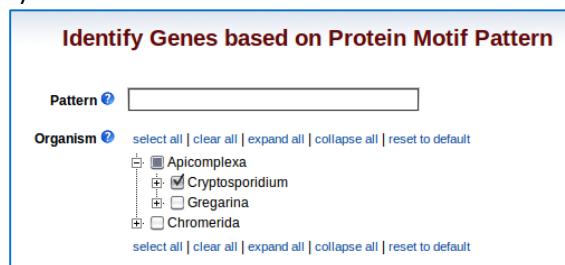
- 1 Intersect 2
- 1 Union 2
- 1 Relative to 2, using genomic colocation
- 1 Minus 2
- 2 Minus 1

A "Run Step" button is located at the bottom of the window. A red arrow points from this button to a final workflow diagram on the right. This diagram shows the "Text 9445 Genes Step 1" box connected to a "Prot Motif 1496 Genes Step 2" box, which is highlighted in yellow. The resulting output is "158 Genes". A red "Add Step" button is also present in this final workflow.

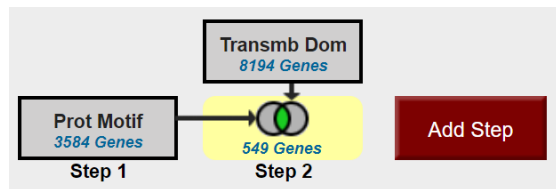
3. Find *Cryptosporidium* genes with the YXXΦ receptor signal motif. Note: for this exercise use <http://cryptodb.org>

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. \*\*\*Note: do not look for the Φ symbol on your keyboard – this will not work. Rather you should use the amino acid symbols.

- 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.

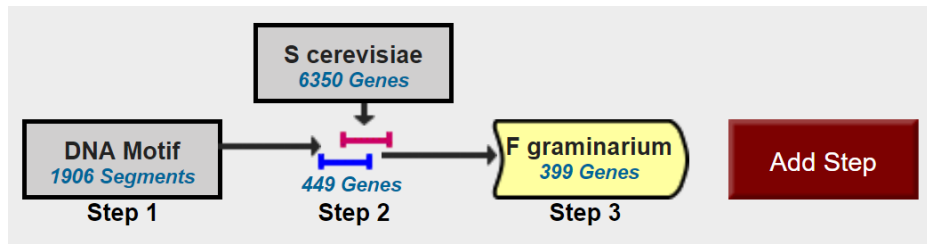


- c. What would happen if you revise the first step (the motif pattern step) to include genes with the sorting motif in the C-terminal 20 amino acids? (hint: edit the first step and modify your regular expression) <http://cryptodb.org/cryptodb/im.do?s=37e8b03ea8087b5a>

4. Find fungal genes downstream of a regulatory DNA motif. For this exercise use: <http://fungidb.org>

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

The goal of this exercise is to identify all genes harboring upstream CACGTG motif, known for its role in transcriptional regulation. We will start our search in an extensively studied model organism *Saccharomyces cerevisiae*, and expand our search to *Fusarium graminearum*. Here is a summary of the search strategy:



**a. Find the CACGTG DNA motif in the *Saccharomyces cerevisiae* genome.**

- Select the “Search for genomic segments (DNA motif)” menu from the Search menu and look for CACGTG in *S. cerevisiae*.

**Identify Genomic Segments based on DNA Motif Pattern**

The screenshot shows a search interface. On the left, a search box is labeled 'DNA' and 'Genomic Segments - DNA Motif Pattern'. On the right, a taxonomic tree is displayed with 'Saccharomyces cerevisiae S288c' selected. Below the tree, a 'Pattern' field contains 'CACGTG' and a 'Get Answer' button is at the bottom.



- Your search returns over 1900 DNA segments containing GACGTG motif. Next, let’s look for putative regulatory targets of this motif by searching for genes that are located 600bp downstream of this sequence.

**b. Identify genes with the CACGTG motif located 600bp upstream of an open reading frame.**

EuPathDB offers a colocation function to identify genomic features within a specified distance of each other. Run a search for all genes in *Saccharomyces cerevisiae* and use the colocation tool to identify genes that contain the CACGTG motif in their upstream regions

- Click “Add Step”. Choose “Run a new search for Genes” > “Taxonomy” > “Organism” and select “Relative to genomic location”.
- Set up the colocation using the following guidelines:


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.

**Genomic Colocation**  

Combine Step 1 and Step 2 using relative locations in the genome  
You had **1906 Genomic Segments** in your Strategy (Step 1). Your new **Genes** search (Step 2) returned **6918 Genes**.

"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**"

(6918 Genes in Step)



Region  
Gene

Exact

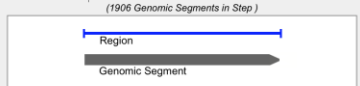
Upstream: 600 bp

Downstream: 1000 bp

Custom:

begin at: start - 600 bp  
end at: start - 1 bp

(1906 Genomic Segments in Step)



Region  
Genomic Segment

Exact

Upstream: 1000 bp

Downstream: 1000 bp

Custom:

begin at: start + 0 bp  
end at: stop + 0 bp

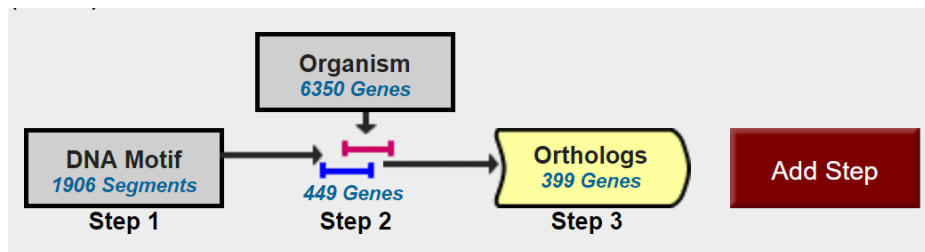
Submit

**c. Identify orthologs *S. cerevisiae* genes in *Fusarium graminearum*.**

All EuPathDB sites offer “Transform by Orthology” function, which is a comparative genomics approach to identify gene orthologs.

This function uses known classifications, OrthoMCL algorithm, and BLAST similarity search to order protein-coding genes from available sequenced genomes into groups of orthologs based on their similarity across multiple species.

Use “Add step” to initiate transformation by orthology into *F. graminearum*.



**d. Investigate GO enrichment and Metabolic pathways records via Analyze results tab.**

**Gene Ontology** features three structured ontologies that describe gene products in terms of their (1) associated biological processes, (2) cellular components role, and (2) molecular functions in a *species-independent* manner.

Biological Process includes processes like the cell cycle, DNA replication, limb formation, etc. Cellular Component assigns gene function to location (for example, a gene product can function in an organelle and/or be a functional component of an enzyme complex). Molecular Function deals with the function/s carried out by a gene product.

It is not uncommon for the same gene product, especially if it belongs to multi-protein enzyme complex, to carry out multiple functions. In this case the same gene product may be identified in several molecular function that make up a biological process.

- Explore GO enrichment for cellular component via **Analyze Results** tab.
- How many GO terms were enriched in the Biological Process search?
- What enriched GO term has the highest statistical significance?
- Based on the information, what function does the group of genes in the result serve?

**Metabolic Enrichment function:** Visualization of enzymatic and chemical flows within biosynthetic pathways.

Initiate a Metabolic Pathway enrichment from the Analyze Results tab.

- How many metabolic processes were identified in your search?
- Which pathway has the best p-values?
- Visit the pathway record pages for ec00240 Pyrimidine metabolism and explore the pathway