

## Motif Searches and Regular Expressions Exercise 9

### 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. (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?

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** select all | clear all

- 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

Get Answer

(Genes)

Text  
14987 Genes

Add S

Step 1

Add Step

Run a new Search for

Transform by Orthology

Add contents of Basket

Add existing Strategy

Filter by assigned Weight

Genes ←

Genomic Segments (DNA)

Motif

SNPs

ORFs

SAGE Tags

Text, IDs, Species

Genomic Position

Gene Attributes

Protein Attributes

Protein Features ←

Similarity/Pattern ←

Transcript Expression

Protein Expression

Cellular Location

Putative Function

Evolution

Protein Motif Pattern

Interpro Domain ←

BLAST

Close

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** PFAM

**Specific Domain(s)**

Begin Or on

PF00920 : Ded\_cyto Deducator of cytokinesis

PF05804 : KAP Kinesin-associated protein (KAP)

PF00225 : Kinesin Kinesin motor domain

Advanced Parameters

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

1 Intersect 2

1 Minus 2

1 Union 2

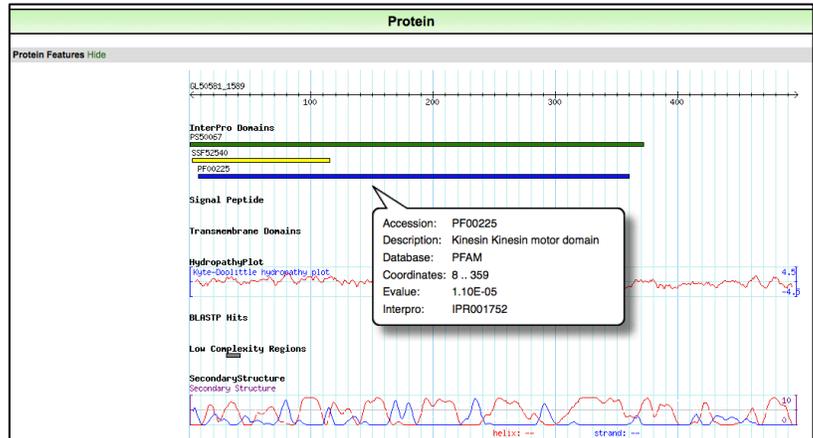
2 Minus 1

1 Relative to 2, using genomic colocation

Run Step

(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.)

- 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 LbrM.32.0490 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.)



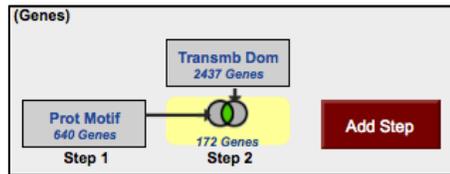
## 6.2 Using regular expressions to find motifs in CryptoDB: finding genes with the YXXΦ receptor signal motif

- a. 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.
- b. 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).

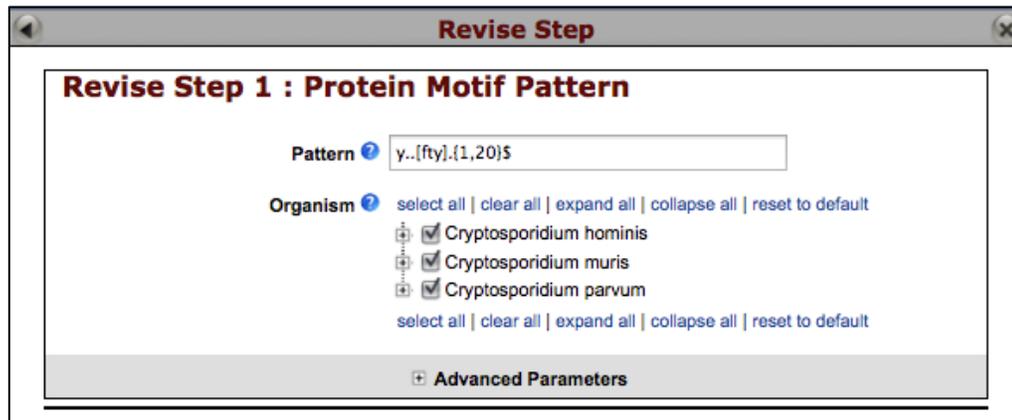
The screenshot shows the search interface for identifying genes based on a protein motif pattern. The interface includes:

- Pattern:** A text input field for entering the regular expression.
- Organism:** A list of organisms with checkboxes:  Cryptosporidium hominis,  Cryptosporidium muris, and  Cryptosporidium parvum.
- Advanced Parameters:** A button to expand search options.
- Get Answer:** A button to execute the search.

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



d. 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=f8b92af87d10013f>