

Motif Searches and Regular Expressions

Exercise 5

5.1 Using InterPro domain searches to identify unannotated kinesin motor proteins.

For this exercise use <http://tritrypdb.org>

a. Identify all genes annotated as hypothetical in *L. braziliensis*.

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

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

Organism select all | clear all | expand all | collapse all | reset to default

- Leishmania
 - Leishmania braziliensis
 - Leishmania infantum
 - Leishmania major
 - Leishmania mexicana
 - Leishmania tarentolae
- Trypanosoma

select all | clear all | expand all | collapse all | reset to default

Text term (use * as wildcard)

Fields select all | clear all

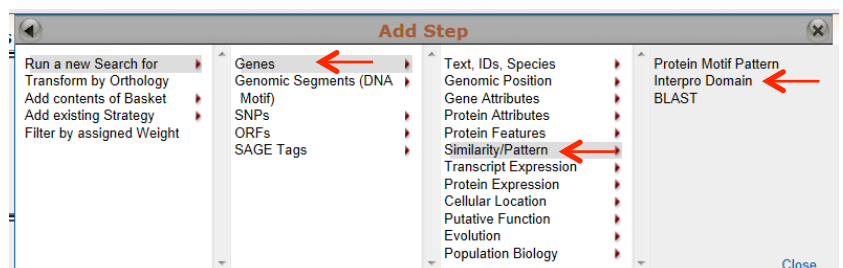
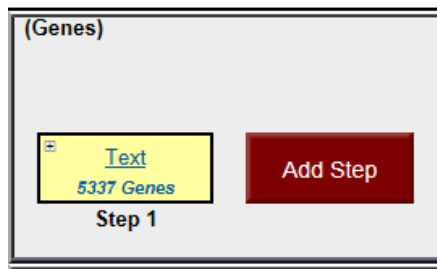
- Gene ID
- Alias
- Gene product
- Phenotype
- GO terms and definitions
- Gene notes
- User comments
- Protein domain names and descriptions
- Similar proteins (BLAST hits v. NRDB/PDB)
- EC descriptions

select all | clear all

Advanced Parameters
 Give this search a weight
 Give this search a name

b. How many of these hypothetical genes have a kinesin-motor protein InterPro 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.



Add Step

Add Step 2 : Interpro Domain

Organism [select all](#) | [clear all](#) | [expand all](#) | [collapse all](#) | [reset to default](#)

- Leishmania
 - Leishmania braziliensis
 - Leishmania infantum
 - Leishmania major
 - Leishmania mexicana
 - Leishmania tarentolae
- Trypanosoma

[select all](#) | [clear all](#) | [expand all](#) | [collapse all](#) | [reset to default](#)

Domain Database **INTERPRO**

Domain

- IPR009079 : 4_helix_cytokine-like_core
- IPR001811 : Chemokine_IL8
- IPR008996 : Cytokine_IL1-like
- IPR001752 : Kinesin_motor_dom
- IPR020091 : Midkine_heparin-bd_GF_dis

Combine Genes in:

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

Run Step

c. Go to the gene page for LbrM.32.0490 and look at the protein feature section. Does this look like a possible motor protein?

Protein

LbrM.32.0490

100 200 300 400 500 600 700

L. braziliensis MS/MS Peptides (Cuervo et. al.)

InterPro Domains

Signal Peptide

Accession: PF00225
 Description: Kinesin Kinesin motor domain
 Database: PFAM
 Coordinates: 23 .. 359
 Value: 1.20E-09
 Interpro: IPR001752

Secondary Structure

Secondary Structure

helix: -- strand: --

BLAST Hits

- hypothetical protein [Leishmania infantum JPCM5]*
- hypothetical protein [Leishmania braziliensis MHOM/BR/75/M2904]*
- hypothetical protein [Leishmania major strain Friedlin]*
- hypothetical protein [Trypanosoma cruzi strain CL Brener]*
- hypothetical protein [Trypanosoma cruzi strain CL Brener]*

5.2 Using regular expressions to find motifs in TriTypDB: finding active trans-sialidases in *T. cruzi*.

- a. *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 1400 genes!!! 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.
 - 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>

