

Sequence Exercises

Motif Searches, Regular Expressions and Genomic 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.

(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 word 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

Get Answer

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

Close

(Genes)

Text

14987 Genes

Step 1

Add Step

Add Step

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)

PF06920 : Ded_cyto Dedicator 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

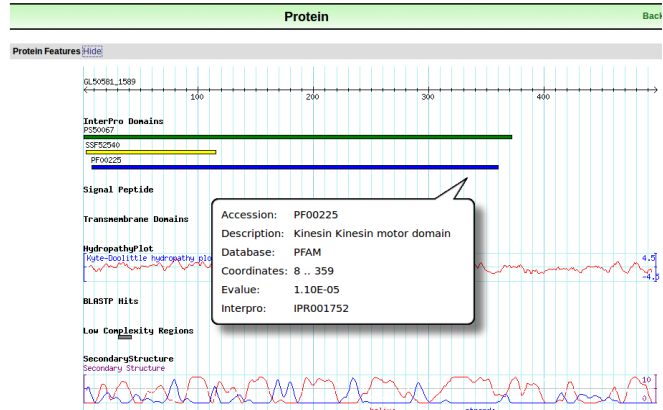
1 Union 2 2 Minus 1

1 Relative to 2, using genomic colocation

Run Step

- 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>

- 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 3500 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.

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.

Add Step 2 : Protein Motif Pattern

Pattern

Organism

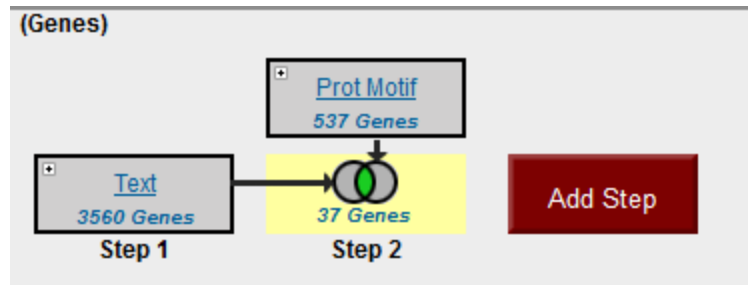
- ☐ Leishmania
- ☒ Trypanosoma
 - ☐ Trypanosoma brucei
 - ☐ Trypanosoma congolense
 - ☒ Trypanosoma cruzi
 - ☐ Trypanosoma evansi
 - ☐ Trypanosoma vivax

☐ Advanced Parameters

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

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.
- 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).

Identify Genes based on Protein Motif Pattern

Pattern 

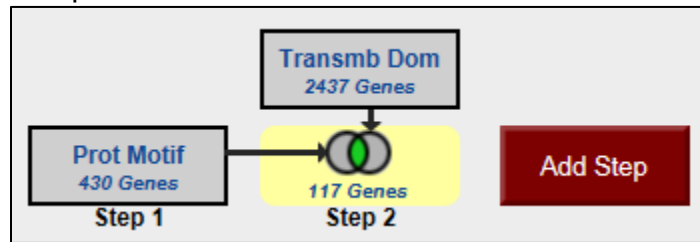
Organism  [select all](#) | [clear all](#) | [expand all](#) | [collapse all](#) | [reset to default](#)

- ☒ *Cryptosporidium hominis*
- ☒ *Cryptosporidium muris*
- ☒ *Cryptosporidium parvum*

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

[Advanced Parameters](#)

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

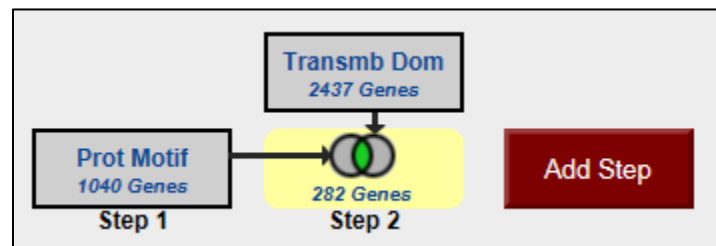


d. What would happen if you revise the first step (the motif pattern step) to include genes with the sorting motif in the C-terminal 20 amino acids? (hint: edit the first step and modify your regular expression).

The screenshot shows the 'Revise Step' interface for Step 1: Protein Motif Pattern. The 'Pattern' field contains the regular expression 'y:[ftY]{0,16}\$'. The 'Organism' section lists three organisms: Cryptosporidium hominis, Cryptosporidium muris, and Cryptosporidium parvum, all of which are selected. Below the organisms is an 'Advanced Parameters' section. A 'Run Step' button is at the bottom.

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

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



4. Identification of specific DNA motifs.

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

- Find all *Bam*HI restriction sites in all microsporidia genomic sequences available in MicrosporidiaDB. Note: you can use the DNA motif search to find complex motifs like transcription factor binding sites using regular expressions.

Hint: *Bam*HI = GGATCC and the DNA motif search is under the heading “Genomic Segments”.

Identify Genomic Segments based on DNA Motif Pattern

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

- ☒ Nosema
- ☒ Encephalitozoon
- ☒ Spraguea
- ☒ Edhazardia
- ☒ Anncalia
- ☒ Enterocytozoon
- ☒ Hamiltosporidium
- ☒ Vavraia
- ☒ Vittiforma
- ☒ Nematocida

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

Pattern GGATCC

Advanced Parameters

- How many times does the *Bam*HI site occur in the genomes you searched? Take a look at your results; notice the Genomic location and the Motif columns.

My Strategies: [New](#) [Opened \(1\)](#) [All \(1\)](#) [Basket](#) [Public Strategies \(5\)](#) [Help](#)

(Segments) Strategy: DNA Motif *

[DNA Motif](#) 27206 Segments [Add Step](#)

Step 1

27206 Genomic Segments from Step 1 [Add 27206 Genomic Segments to Basket](#) | [Download 27206 Genomic Segments](#)

Strategy: DNA Motif

Genomic Segment Results Genomic Locations

First 1 2 3 4 5 Next Last [Advanced Paging](#) [Add Columns](#)

Segment ID	Organism	Genomic Location	Motif
KK358017:490-496.f	Anncalia algerae PRA109	KK358017: 490 - 496 (+)	...ATATATTGAAGCAAATTTATGGATCCGCTGTATCCTTAAAGTCGA...
KK358017:490-496.r	Anncalia algerae PRA109	KK358017: 490 - 496 (-)	...TCGACTTTAAGGATAACAGCGGATCCATAAATTTGCTTCAATATAT...
KK358017:6265-6271.f	Anncalia algerae PRA109	KK358017: 6265 - 6271 (+)	...TAATATCTTGGATCATTGGATCCTGAITTTGGTACTTTATTAA...

c. Find genes that have one of these *Bam*HI sites within 500 nucleotides upstream of their start.

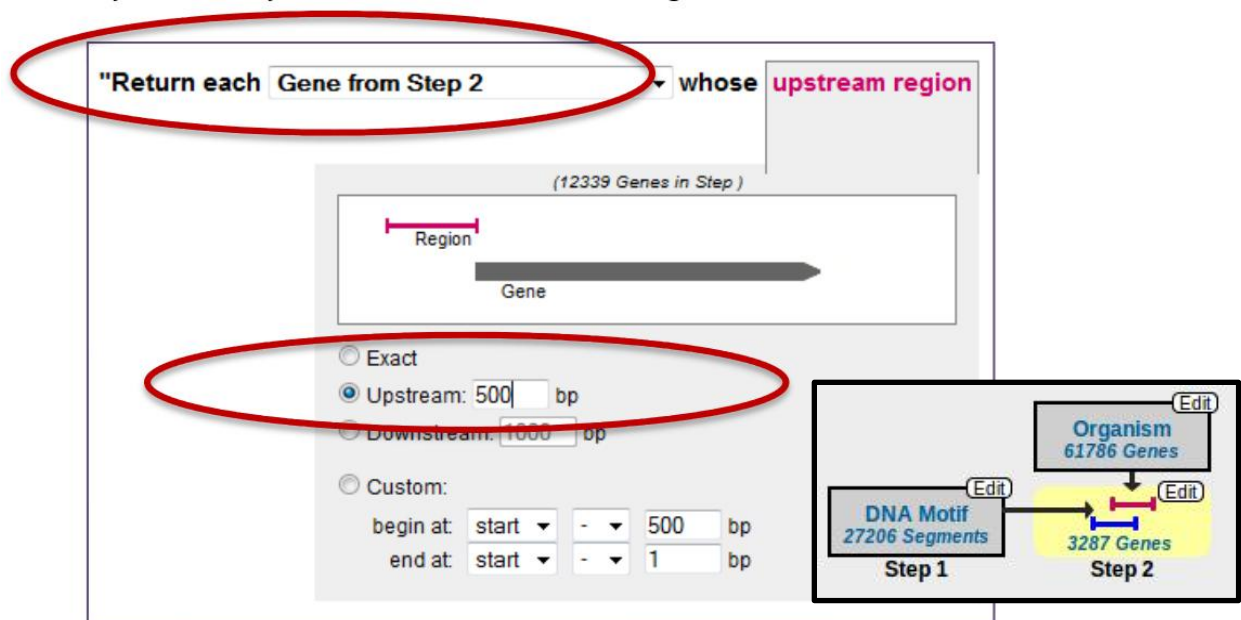
In section 1 you found *Bam*HI sites, but now you are looking for genes that have one of these sites located within 500 nucleotides upstream of their start.

Hint: You can achieve this by running a genomic collocation search that defines the genomic relationship between the *Bam*HI sites and genes. Add a “Genes by Organism” step to the motif search and select the “1 relative to 2, using genomic locations” option.

The screenshot shows the Genomic Collocation search interface. It includes a 'My Strategies' panel on the left with a 'DNA Motif' strategy. The main area shows a list of genomic segments. The 'Add Step' dialog is open, showing a list of organisms. The 'Add Step 2: Organism' dialog is also open, showing a list of organisms. The 'Combine Genomic Segments in Step 1 with Genes in Step 2' section is visible, showing options for '1 Relative to 2, using genomic collocation'.

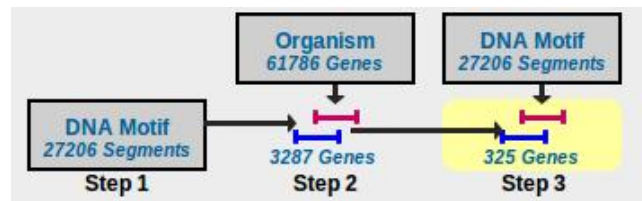
The screenshot shows the 'Genomic Collocation' search results and configuration panel. It includes a 'Return each Gene from Step 2' dropdown and a 'whose upstream region overlaps the exact region of a Genomic Segment in Step 1' dropdown. The configuration panel shows options for 'Exact', 'Upstream', and 'Downstream' regions. The 'Upstream' region is selected, with a value of 500 bp. The 'Exact' region is selected, with a value of 1000 bp. The 'Downstream' region is selected, with a value of 1000 bp. The 'Custom' option is also available, with a 'begin at' and 'end at' field.

How did you modify the location relative to genes? How many genes did you get?

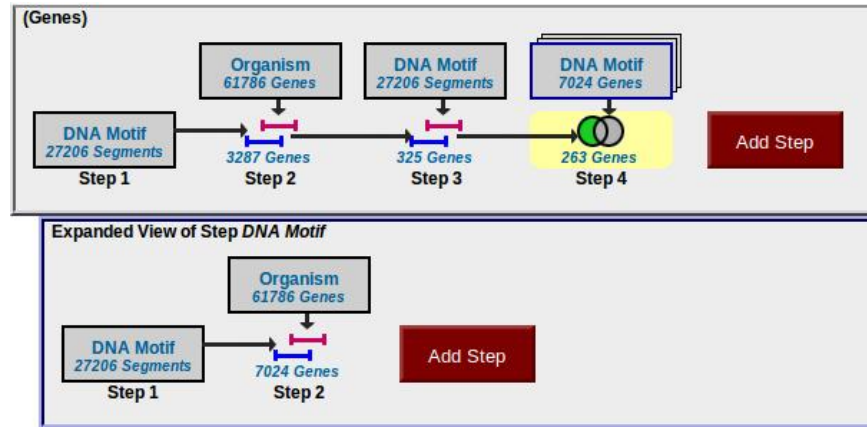


- d. Using a similar sequence of steps as in part 2, define which of these genes also have a *Bam*HI site in their 500 nucleotide downstream region.

Hint: after you click on add step you will have to select DNA motif search and select the genomic collocation option.



- e. Taking this a step further, define which of these genes do NOT contain a *Bam*HI site within them.



Hint: you will have to use a nested strategy.

Look at your results. Do they make sense? Confirm your results by looking at one of the genes in Gbrowse and showing *Bam*HI restriction sites.

Note: you can add a column to any result table that allows you to go directly to GBrowse at the genomic coordinates of any ID in your result list. Click on the Add Columns button.

The screenshot displays the Gbrowse interface with a table of 263 genes from Step 4. The table includes columns for Gene ID, Genomic Location, and Protein Name. A 'Select Columns' dialog box is open, allowing users to customize the table columns. The dialog box has a list of available columns and a 'Select Columns' button. The 'Add Columns' button is highlighted with a red circle and an arrow pointing to it.

Table Data (263 Genes from Step 4):

Gene ID	Genomic Location	Protein Name
EBI_24411	ABGB01000099: 438 - 728 (+)	hypothetical protein
EBI_27581	ABGB01000203: 976 - 1,491 (-)	hypothetical protein
EBI_25435	ABGB01000276: 1,036 - 1,248 (-)	hypothetical protein
EBI_26304	ABGB01000351: 1,323 - 1,454 (+)	hypothetical protein
EBI_26621	ABGB01000486: 358 - 558 (+)	hypothetical protein
EBI_25638	ABGB01000541: 218 - 430 (-)	hypothetical protein
EBI_25705	ABGB01000850: 191 - 403 (+)	hypothetical protein
EBI_26491	ABGB01000853: 329 - 541 (-)	hypothetical protein
EBI_26598	ABGB01000992: 532 - 744 (+)	hypothetical protein
EBI_27558	ABGB01001170: 475 - 687 (+)	hypothetical protein
EBI_27632	ABGB01001257: 59 - 238 (+)	aspartate aminotransferase
EBI_25657	ABGB01001308: 181 - 393 (+)	hypothetical protein

Select Columns Dialog Box:

- Text, IDs, Species:** ☐ Text, IDs, Species
- Genomic Position:** ☐ Chromosome, ☒ Genomic Location, ☐ Gene Strand
- Gene Attributes:** ☐ Gene Attributes
- Protein Attributes:** ☒ Product Description, ☐ Molecular Weight, ☐ Isoelectric Point
- Protein Features:** ☐ Protein Features
- Transcript Expression:** ☐ Transcript Expression
- Putative Function:** ☐ Putative Function
- Evolution:** ☐ Evolution
- Search PDB by the protein sequence:** ☐ Search PDB by the protein sequence
- GBrowse:** ☒ GBrowse
- Weight:** ☐ Weight

The 'Add Columns' button is highlighted with a red circle and an arrow pointing to it.

Note: you can configure restriction sites by clicking on the configure button in GBrowse and selecting the restriction sites you would like to display. To view restriction sites, the “Restriction Sites” data track must be turned on. Go to the “Select Tracks” page and click “Restriction Sites” under the “Analysis” section.

Browser | Select Tracks | Snapshots | Custom Tracks | Preferences

Search

Landmark or Region: NC_003229:162,593..182,592

Annotate Restriction Sites

Scroll/Zoom: << < - Show 20 kbp + > >> ☐ Flip

Overview

Region

Details

NC_003229: 20 kbp

5 kbp

★ ■ ■ ■ ■ Annotated Genes (with UTRs in gray when available)

ECU02_1360 ECU02_1370 ECU02_1380 ECU02_1390 ECU02_1400 ECU02_1410 ECU02_1420 ECU02_1430 ECU02_1440

Select Tracks Clear highlighting

The restriction site plugin generates a restriction map on the current view.
This plugin was written Elizabeth Nickerson & Lincoln Stein.

Select Restriction Sites To Annotate

Restriction Site Display ☐ off ☒ on

<input type="checkbox"/> AatII	<input type="checkbox"/> BspDI	<input type="checkbox"/> HpaII	<input type="checkbox"/> PspGI
<input type="checkbox"/> Acc65I	<input type="checkbox"/> BspEI	<input type="checkbox"/> Hpy188I	<input type="checkbox"/> PspOMI
<input type="checkbox"/> AccI	<input type="checkbox"/> BspHI	<input type="checkbox"/> Hpy188III	<input type="checkbox"/> PstI
<input type="checkbox"/> AcII	<input type="checkbox"/> BsrFI	<input type="checkbox"/> Hpy99I	<input type="checkbox"/> Pvul
<input type="checkbox"/> Afel	<input type="checkbox"/> BsrGI	<input type="checkbox"/> HpyCH4III	<input checked="" type="checkbox"/> PvuII
<input type="checkbox"/> AffII	<input type="checkbox"/> BssHII	<input type="checkbox"/> HpyCH4IV	<input type="checkbox"/> RsaI
<input type="checkbox"/> AflIII	<input type="checkbox"/> BssKI	<input type="checkbox"/> HpyCH4V	<input type="checkbox"/> RsrII
<input type="checkbox"/> AgeI	<input type="checkbox"/> BstAPI	<input type="checkbox"/> KasI	<input type="checkbox"/> SacI
<input type="checkbox"/> AhdI	<input type="checkbox"/> BstBI	<input type="checkbox"/> KpnI	<input type="checkbox"/> SacII
<input type="checkbox"/> AluI	<input type="checkbox"/> BstEII	<input type="checkbox"/> MboI	<input type="checkbox"/> SalI
<input type="checkbox"/> AlwNI	<input type="checkbox"/> BstNI	<input type="checkbox"/> MfeI	<input type="checkbox"/> Sau3AI
<input type="checkbox"/> ApaI	<input type="checkbox"/> BstUI	<input type="checkbox"/> MluI	<input type="checkbox"/> Sau96I
<input type="checkbox"/> ApaLI	<input type="checkbox"/> BstXI	<input type="checkbox"/> MscI	<input type="checkbox"/> SbfI
<input type="checkbox"/> ApoI	<input type="checkbox"/> BstYI	<input type="checkbox"/> MseI	<input type="checkbox"/> Scal
<input type="checkbox"/> AscI	<input type="checkbox"/> BstZ17I	<input type="checkbox"/> MslI	<input type="checkbox"/> ScrFI
<input type="checkbox"/> AseI	<input type="checkbox"/> Bsu36I	<input type="checkbox"/> MspA1I	<input type="checkbox"/> SexAI

Overview

Region

Details

AL598442: 30.81 kbp

10 kbp

★ ■ ■ ■ ■ Restriction Sites

BamHI restriction site

BamHI BamHI BamHI BamHI

★ ■ ■ ■ ■ Annotated Genes (with UTRs in gray when available)

ECU02_1310 ECU02_1340 ECU02_1370 ECU02_1390 ECU02_1410 ECU02_1440 ECU02_1460 ECU02_1480 ECU02_1500 ECU02_1530 ECU02_1550

ECU02_1320 ECU02_1350 ECU02_1380 ECU02_1400 ECU02_1420 ECU02_1450 ECU02_1470 ECU02_1490 ECU02_1510 ECU02_1520