

# 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 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 cyto kinesins  
 PF05804 : KAP Kinesin-associated protein (KAP)  
 PF00225 : Kinesin Kinesin motor domain

Free Text (use "\*" for wildcard)

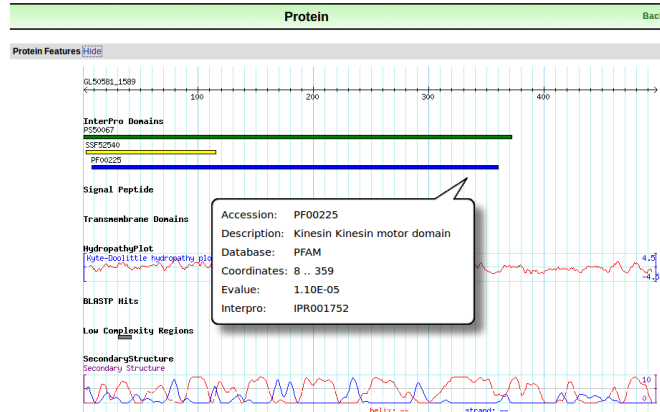
Advanced Parameters

**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 CryptoDB: finding genes with the YXXΦ receptor signal motif

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

- 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).

### Identify Genes based on Protein Motif Pattern

Pattern

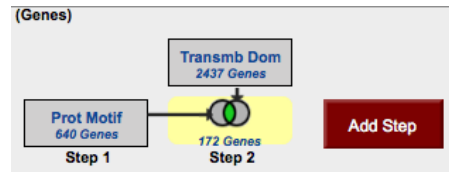
Organism  select all | clear all | expand all | collapse all | reset to default

- 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.

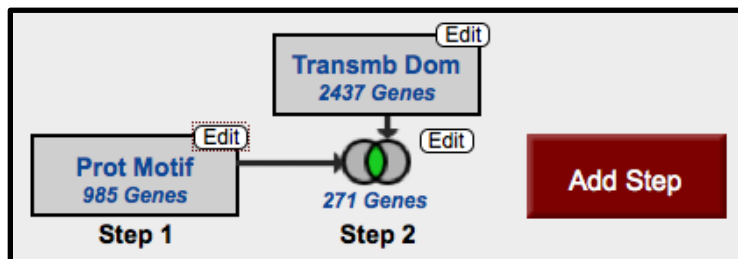


d. What would happen if you revise the first step (the motif pattern step) to include genes with the sorting motif in the C-terminal 20 amino acids? (hint: edit the first step and modify your regular expression).

The screenshot shows a window titled 'Revise Step'. Inside, the title is 'Revise Step 1 : Protein Motif Pattern'. There is a 'Pattern' field with the value 'y.[fyt]{0,16}\$'. Below it, the 'Organism' section is expanded, showing three checked options: 'Cryptosporidium hominis', 'Cryptosporidium muris', and 'Cryptosporidium parvum'. At the bottom of the dialog is an 'Advanced Parameters' section. Below the dialog is a 'Run Step' button.

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

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



### 3. Identification of specific DNA motifs.

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

- a. 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”.

The screenshot shows the MicrosporidiaDB search interface. On the left, there are three main sections: 'Identify Genes by:', 'Identify Other Data Types:', and 'Tools:'. The 'Identify Other Data Types:' section is expanded to show 'Genomic Segments (DNA Motif)'. A red arrow points from this section to a pop-up window titled 'Identify Genomic Segments based on DNA Motif Pattern'. This window contains a list of organisms with checkboxes, a 'Pattern' input field containing 'GGATCC', and a 'Get Answer' button.

- b. 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.

The screenshot shows the search results page for the DNA motif GGATCC. The page title is '27206 Genomic Segments from Step 1'. Below the title, there is a table with the following columns: Segment ID, Organism, Genomic Location, and Motif. The table contains three rows of results.

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 (-)	...TCGACITTAAGGATAACAGCGGATCCATAAAATTTGCTCAATATAT...
KK358017:6265-6271.f	Anncalia algerae PRA109	KK358017: 6265 - 6271 (+)	...TAATTATTCTGGATCATTCCGGATCCGTATTGGTACTTTATTAA...

- 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 'My Strategies' interface. A red arrow labeled '1' points to the 'Add Step' button. A second red arrow labeled '2' points to the 'Add Step' dialog box, which has 'Genes' selected. A third red arrow labeled '3' points to the 'Text, IDs, Organism' option in the 'Add Step' dialog. A fourth red arrow labeled '4' points to the '1 Relative to 2, using genomic collocation' option in the 'Combine Genomic Segments' section.

5

**Genomic Collocation** ?

Combine Step 1 and Step 2 using relative locations in the genome  
 You had 27206 Genomic Segments in your Strategy (Step 1). Your new Genes search (Step 2) returned 61786 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"

The diagram shows two panels. The left panel is for 'Gene from Step 2' and has 'Upstream: 500 bp' selected. The right panel is for 'Genomic Segment in Step 1' and has 'Exact' selected. A central diagram shows a red arrow pointing from the gene's upstream region to the genomic segment's exact region.

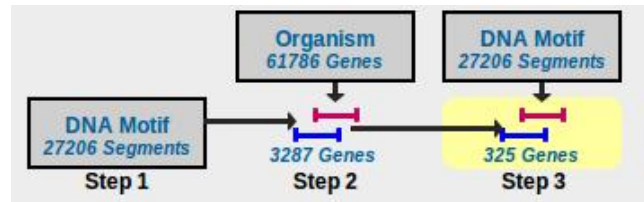
Submit

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

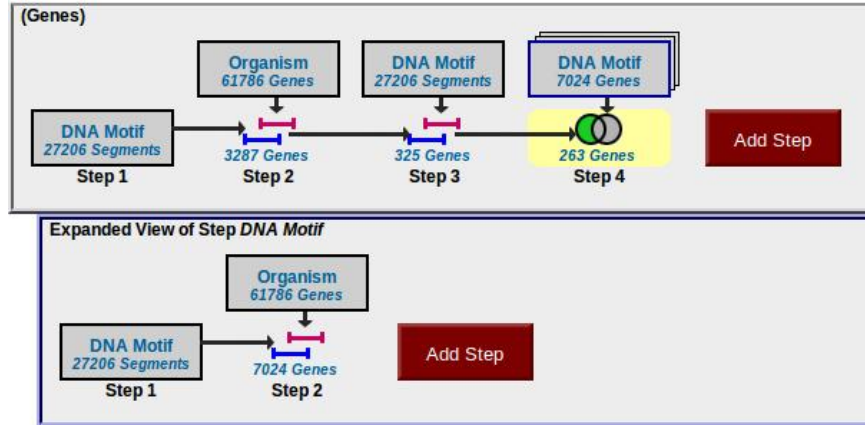
The screenshot shows a bioinformatics tool interface. At the top, a dropdown menu is set to "Return each Gene from Step 2" and the filter is set to "whose upstream region". Below this, a diagram shows a "Region" (pink bar) above a "Gene" (grey arrow). The search criteria are:   
-  Exact   
-  Upstream: 500 bp   
-  Downstream: 1000 bp   
-  Custom: begin at start - 500 bp, end at start - 1 bp   
A workflow diagram on the right shows:   
- Step 1: DNA Motif (27206 Segments)   
- Step 2: 3287 Genes (highlighted in yellow)   
- Organism (61786 Genes)   
Arrows indicate the flow from Step 1 to Step 2, and from Organism to Step 2.

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.

263 Genes from Step 4  
Strategy: DNA Motif

Add 263 Genes to Basket | Download 263 Genes

Click on a number in this table to limit/filter your results

All Results	Ortholog Groups	Anncalia		Edhazardia		E.cuniculi (nr Genes: 37)		
		A.algerae (nr Genes: 10)	E.aedis	PRA109	PRA339	USNM 41457	EC1	EC2
263	160	5	5	0	35	32	32	

Gene Results | Genome View

First 1 2 3 4 5 Next Last | Advanced Paging

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

Select Columns

Update Columns

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

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

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

Update Columns

Nematocida  
(nr : 2) | N.sp. 1 (nr Genes: 6)  
ERTm3 | ERTm2 | ERTm6  
1 | 3 | 3

Add Columns

**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  Search

Annotate Restriction Sites

Save Snapshot

Data Source: MicrosporidiaDB GBrowse v2.48

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

Overview

NC\_003229

Region

Details

NC\_003229: 20 kbp

5 kbp

163k 164k 165k 166k 167k 168k 169k 170k 171k 172k 173k 174k

★      Annotated Genes (with UTRs in gray when available)

ECU02\_1360 ECU02\_1380 ECU02\_1400 ECU02\_1420 ECU02\_1440

ECU02\_1370 ECU02\_1390 ECU02\_1410 ECU02\_1430

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"/> AclI	<input type="checkbox"/> BsrFI	<input type="checkbox"/> Hpy99I	<input type="checkbox"/> PvuI
<input type="checkbox"/> AfeI	<input type="checkbox"/> BsrGI	<input type="checkbox"/> HpyCH4III	<input checked="" type="checkbox"/> PvuII
<input type="checkbox"/> AflI	<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"/> Sall
<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

AL590442

Region

Details

AL590442: 30.81 kbp

10 kbp

160k 170k 180k

★      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\_1320 ECU02\_1350 ECU02\_1380 ECU02\_1400 ECU02\_1420 ECU02\_1450 ECU02\_1470 ECU02\_1500 ECU02\_1530

ECU02\_1330 ECU02\_1360 ECU02\_1430 ECU02\_1490 ECU02\_1520