Motif Searches and Regular Expressions Exercise 6

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

Note: 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

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.

Identify Genes base	ed on Text (product name, notes, etc.)
Organism 🛛	select al clear al expand al collapse al reset to default
Text term (use * as wildcard) 📀	hypothetical
Fields 🛛	Gene ID Alias Gene product Fhenotype 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
	Get Answer



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? Hint: click on the ID for LbrM.32.0490 in the result table to go to the gene page. Mouse over the glyphs in the Protein Features graphic.



6.2 Using regular expressions to find motifs in TriTypDB: finding active transsialidases 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 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 transsialidases 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 transsialidases.
 - 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 <u>regular</u> <u>expression tutorial</u> if you need to.

.	Add Step 2 : Protein Motif Patter	n
	Pattern 😢	^m.{340}y
	Organism 🕖	select all clear all expand all collapse all reset to default
		🕂 🖂 Leishmania
		Trypanosoma brucei
		Trypanosoma congolense
·		🕀 🗹 Trypanosoma cruzi
		Trypanosoma evansi
		T: Trypanosoma vivax
		celect all clear all expand all collance all recet to default
		select all ficear all expand all collapse all reset to default
•	. ∎ Ad	Ivanced Parameters
-	Combine Genes in Step 1 with Ge	enes in Step 2:
	~	~
	O 1 Interse O	ct 2 💿 🚺 1 Minus 2
	① 1 Union 2	O 2 Minus 1 O
	🔘 💾 1 Relative	e to 2, using genomic colocation
		Run Step

If you need help, you can go to this sample strategy below to see the answer: <u>http://tritrypdb.org/tritrypdb/im.do?s=a905e36f634f7b42</u>



6.3 Identification of specific DNA motifs. Note: For this exercise use <u>http://microsporidiadb.org</u>

a. Find all BamHI 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: BamHI = GGATCC and the DNA motif search is under the heading "Genomic Segments".



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

My Stra	tegies: New	Opened (1)	All (1) 🗇 Basket	Examples	Help			
(Segments)					Strategy:	DNA Motif *	× ^
DNA	Motif						Rename Duplicate Save As Share Delete	e s e
20628 Se Step	Add Step							-
			3	-				
20628 C Strategy Genomi	Genomic Segments y: DNA Motif c Segment Results 2 3 4 5 Next Last	from Step 1 Genomic Locations Advanced P	Add 20628 (aging	Genomic Segm	ents to Ba	sket Download 20628 G	enomic Segn Add Colum	nents
\$ s	egment ID	韋 Organism 🕹	Genomic Location	🕲 🌲 Motif 🕲)			
CAIR	01000013.1:1604-1610:f	Anncaliia algerae Undeen	CAIR01000013.1: 1604 - 1610 (+)	AAACAA	AGTITACAA	CAGTGGGATCCAATCACTG	TTCCTCCGACA	4C
	01000013.1:1604-1610:r	Anncaliia algerae Undeen	CAIR01000013.1: 1604 - 1610 (-)	GTGTCG	GAGGAACA	GTGATTGGATCCCACTGTTC	GTAAACTTTGTT	т
CAIR	01000037.1:501-507:f	Anncaliia algerae Undeen	CAIR01000037.1: 501 - 507 (+)	TTATTATT	TATGCATTG	AATGGATCCCTTTTTGCATA	AATTAAAAA	
CAIR	01000037.1:501-507:r	Anncaliia algerae Undeen	CAIR01000037.1: 501 - 507 (-)	TTTTTAA	TTTATGCAA	AAAG <mark>GGATCC</mark> ATTCAATGCA	TAAATAATAA	
CAIR	01000050.1:666-672:f	Anncaliia algerae Undeen	CAIR01000050.1: 666 - 672 (+)	TTGTGTG	GACGCCTC	GTGTCA <mark>GGATCC</mark> TTGAAAAA	TTTTGAGTGAT	Т

6.4 Find genes that have one of these BamHI sites within 500 nucleotides upstream of their start.

In the section 7.3 you found BamHI 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 BamHI sites and genes. Add a "Genes by Organism" step to the motif search and select the "1 relative to 2, using genomic locations" option.



			Add Step			×
5		G	enomic Colocatio	n የ 🗣		
J		Combine Step 1 and	d Step 2 using relative	locations in	the genome	
	You had 2	0628 Genomic Segments in your	Strategy (Step 1). Your n	ew Genes sear	ch (Step 2) returned 35231 Genes.	
	a (a a			· · ·		
Return each	Gene from Step 2	 whose upstream region 	overlaps + the	exact region	of a Genomic Segment in Step 1 and is on either stra	nd 🔻
	(35231	Genes in Step)			(20628 Genomic Segments in Step)	
	Region				Region	
	Gene				Genomic Segment	
	© Exact			Exact		
	Opstream: 500 bp			O Upstream:	1000 bp	
	© Downstream: 1000 bp			O Downstream	m: 1000 bp	
	Custom:			Custom:		
	begin at: start 🔻 -	▼ 500 bp		begin at:	start 🔻 + 💌 0 bp	
	end at: start 👻 -	▼ 1 bp		end at:	stop 🔻 + 👻 0 bp	
			Submit			
						Close

How did you modify the location relative to genes?

"Return each Ge	ne from Step 2 whose upstream region
	(12339 Genes in Step)
	Region
	Gene
	© Exact
	O Downstream. 1999 bp
	Custom:
	begin at: start ▼ - ▼ 500 bp end at: start ▼ - ▼ 1 bp

How many genes did you get? [Genes]



6.5 Using a similar sequence of steps as in part 7.4, define which of these genes also have a BamHI 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.



6.6 Taking this a step further, define which of these genes do <u>NOT</u> contain a BamHI site within them.

Hint: you will have to use a nested strategy.

My Strategies:	New C	pened (1)	All (1)	🔿 Basket	Examples	Help
(Genes) C GGATCC 20628 Segments Step 1	Organism 35231 Genes 2722 Genes Step 2	* DNA Ma 20628 Segr 293 Gen Step 3	es 3	Organism 5490 Genes 232 Genes Step 4	Add Step	
Expanded View of S Organism 35231 Genes Step 1	tep Organism DNA Motif 20628 Segmen 5490 Genes Step 2	Add	d Step			
					_	

Look at your results. Do they make sense? Confirm your results by looking at one of the genes in Gbrowse and showing BamHI 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.

232 G Strate	enes f egy: G	rom St GATCC	tep	4								Add 2	32 Genes to E	asket	Downl	oad 23	2 Genes	;
🗆 🕈 Fi	lter result	s by spe	cies	(resu	its ren	noved	by the fill	ter will no	t be con	nbined into the next	step.)							_
AII	Ortholog	Encep	halito	ozoon	cunic	uli	Ence	ephalitoz hellem	oon	Encophalitazoon	Enconha	litozoon	Entorocutozoon	٨	lematocio	da	Nosoma	
Results	Groups	Distinct genes	EC1	EC2	EC3	GB M1	Distinct genes	ATCC 50504	Swiss	intestinalis	roma	leae	bieneusi	parisii ERTm1	parisii ERTm3	sp. 1 ERTm2	ceranae	
232	106	133	35	32	32	34	21	18	15	23	2	1	12	2	1	3	0	
∢									П	11							Þ	
Gene	Results	Ger	nome	e Viev	v													
First	1234	5 Next I	ast			۵d	vanced	Paging						-	ſ	Add Co	lumns	
	1204		use					r uging								Had Co	i anno	
⊕ ₹	Gene II	D Ģ G	enor	mic L	.ocat	ion	3	₽	roduc	t Description (يل ا							
⊕ EE																		
EL CONTRACTOR	31_27581 81_26436	ABG	B010 B010	10020	13: 97 76: 1 1	'6 - 1 136	,491 (-)	hypo	othetica	al protein				_ J	/			
	51_25455)	DUTU	10021	0. 1,0	030 -	1,240 (- nype	uneuca	ai protein	1							
👚 EE	31_26304	ABG 1,45	B010 4 (+)	0035	51: 1,:	323 -		hypo	othetica	al protein		Sele	ect Columns					×
🗁 EE	BI_26621	ABG	B010	00048	36: 35	8 - 5	58 (+)	hypo	thetica	al protein			E	Jodate	Colum	ns		
🗁 EE	BI_25638	ABG	B010	00054	1: 21	8 - 4	30 (-)	hypo	othetica	al protein								
⊕ EE	31_25705	ABG	B010	0085	50: 19	1 - 4	03 (+)	hypo	othetica	al protein			clear all reset to	current	d all co reset t	ollapse a to defaul	ll t	
🖀 EE	3I_26491	ABG	B010	00085	53: 32	9 - 5	41 (-)	hypo	othetica	al protein		÷.	Text IDs Sn	ecies	1.00011			
tan EE	3I_26598	ABG	B010	00099)2: 53	2 - 7	44 (+)	hypo	othetica	al protein			Genomic Po:	sition				
1 EE	31_27558	ABG	B010	0111/	0:4/	5-6	87 (+)	hypo	othetica	al proteín			- Chromoso	ome				
	31_27632	ABG	B010	0125	o7:59	1 - 23	8(+)	aspa	irtate-a	immonia ligase	_		- 🗹 Genomic	Locatio	n			
	51_25657		BU10	7. 50	004	- 1 - 1 - 1	93 (+) 222 ()	nypo	othetica				📃 🔲 Gene Stra	and				
					~~~			FINI				÷.	Gene Attribu	tes				
												÷ (	Protein Attrib	utes				
													- 🗹 Product D	escripti	on			
													- 📃 Molecular	Weight	t			
													🛄 📃 Isoelectric	: Point				
												÷ (	Protein Feat	ures				
												÷ (	Transcript Ex	cpressio	n			
												Ŧ.	Putative Fun	ction				
												÷.	Evolution				-	
													GBrowso	by the p	protein s	sequenc	e	
													Weight	,	-			
													clear all reset to	current	id all   co :   reset t	ollapse a o defaul	ll t	
													l	Jpdate	Colum	ns		1.

**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 Trac	ks Preferences							
Search											
Landmark or Region: Annotate Restriction Sites Configure Go											
NC_00322	9:162,593182,593	2	Search		Save S	napshot	Load Sna	apshot	-		
Data Source MicrosporidiaDB GBrowse v2.48											
Overview         NC_003229           ok         10k         20k         30k         40k         50k         60k         70k         80k         90k         100k         120k         130k         140k         150k         160k         170k         180k         19k											
Region	Z									'	
Deteile	0k 10k	20k 30k	40k 50k	60k 70k 80k	90k 100k	110k	120k 130k	140k 150k	160k 170k 180	k 190k	
Details	NC_0032	29: 20 kbp	5	kbp						V	
	<del>&lt;++ ++++++</del> 163k 1	64k 165k 16	56k 167k 168k	169k 170k 171	k 172k 17	+ 3k 174	The restricti	on site plugin ge	enerates a restrict	ion map on the cu	
* = × 5 2 1	Annotated Genes ECU02_13	(with UTRs in g	<mark>)ray when availab</mark> CU02-1380	ECU02 1400	ECU02 1420	ECU02 14	This plugin	was written Eliza	abeth Nickerson 8	Lincoln Stein.	
	EC	U02_1370	ECU02_139	0 ECU02_1	410 ECU02	1430	Select Res	striction Sites T	ο Annotate		
							Restriction	Site Display	○ off		
				Select Tracks	Clear biob	lighting	🗖 Aatll	BspDI	🗖 Hpall	🗖 PspGl	
				Coloci Hacita		ingriding	Acc65	🔲 BspEl	Hpy188I	PspOMI	
							Accl	🔲 BspHI	🔲 Hpy188III	Pstl	
							Acll	BsrFI	Hpy99I	Pvul	
							Afel	BsrGl	HpyCH4III	Pvull	
							Afili	BssHll	HpyCH4IV	Rsal	
							Afili	BssKl	HpyCH4V	Rsrll	
							Agel	BstAPI	Kasl	Sacl	
							Ahdl	BstBl	Kpnl	Sacll	
							Alul	BstEll	Mbol	Sall	
								BstNI	Mfel	Sau3Al	
							Apal	BstU	Mlul	Sau96	
							Anal	BetXI	Macl	Sbfl	
								BetVI		Scal	
								Bet717		ScrEl	
							Asci	Bau20			
							Asel	Bsu36l	INISPA11	SexAl	