DNA motifs Exercise 7

7.1 Identification of specific DNA motifs. Note: For this exercise use http://microsporidiadb.org

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".

Microspori	Version 7.0 12 May 14		A EuPathDB Project	
-watrosponas	a Genomics Resource	Gene ID: ECU03_0820i G	Sene Text Search: synth*	
Home New Search - My Str	rategies My Basket (0) Tools	Data Summary Downloads Com	nunity	
Data Summary News and Tweets	The Working with Parasite Databu workshop is June 20th, 2014. To a	ase Resources workshop will be held from October 1 apply and for more details please visit the <u>course de</u>	9-24, 2014 in Hinxton, UK. The deadline for this scription page.	
S May 2014 MicrosporidiaDB 7.0 released 30 January 2014 MicrospondiaDB 6.0 released 14 January 2014 EUPathDB All MicrospoliaBB News >>>	Identify Genes by: Expand All Collapse All Text, IDs, Organism Genomic Position	Identify Other Data Types: Expand All Collapse All Bisolates Genomic Sequences Genomic Sequences (INA Motio	Tools: BLAST Identify Sequence Similarities Sequence Retrieval	
Tweets by @eupathdb Community Resources expand for 13 new items Education and Tutorials	Certer Autoures Protein Features Similarity/Pattern Transcript Expression Cellular Location Putative Function	Containing Control Co	Retrieve Specific Sequences using IDs and coordinates archogen Portal RNA sequence analysis, interactione mass and more egments based on DN	A Motif Pattern
About MicrosporidiaDB	2 Evolution	Organism ♥ select + 60 + 70 + 70	al clear al expand al collapse al reset to de Vosema Encephalitozoon Spraguea Ghazardia Anncalia Enferoovtozoon	sfault
		⊕ @ ⊕ @ ⊕ @ eeect Pattern @ CCATC	Hamiltosporidium Vavraia Nematocida III clear all expand all collapse all reset to de III clear all expand all collapse all reset to de	sfault
			Advanced Parameters	
			Get Answer	Click to open/close th

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.

Ay Strategies: New	Opened (1)	All (1) 🗇 Basket	Examples	Help				
Segments)						Strategy:	DNA Motif*	^
DNA Motif 20628 Segments Step 1							Renam Duplica Save A Shar Dele	ne te As re te
			_					
20628 Genomic Segments	from Step 1							
Genomic Segment Results	Genomic Locations	Add 20628 (Genomic Segm	ents to B	asket Downloa	id 20628 G	Senomic Seg	ments
Strategy: DNA Motif Genomic Segment Results First 1 2 3 4 5 Next Last	Genomic Locations	Add 20628 (Genomic Segm	ents to B	asket Downloa	v 20628 G	Senomic Seg	ments
Strategy: DNA Motif Genomic Segment Results First 1 2 3 4 5 Next Last	Genomic Locations Advanced P	Add 20628 of Paging Cardinal Control C	Genomic Segm	ents to B	asket Downloa	V	Senomic Seg	mns
Strategy: DNA Motif Genomic Segment Results First 1 2 3 4 5 Next Last	Genomic Locations Advanced F Organism Anncaliia algerae Undeen	Add 20628 0 Paging Genomic CAIR01000013.1: 1604 - 1610 (+)	Cenomic Segmo	AGTITACA	ACAGTGGGATCC		Add Colu	mns CAC
Strategy: DNA Motif Genomic Segment Results First 1 2 3 4 5 Next Last	Genomic Locations Advanced P Organism @ Anncaliia algerae Undeen Anncaliia algerae Undeen	Add 20628 0 Paging CAIR01000013.1: 1604 - 1610 (+) CAIR01000013.1: 1604 - 1610 (-)	Cenomic Segme Motif Contract 	AGTTTACA	ACAGTGGGATCC		Add Colu	mns AC
Strategy: DNA Motif Genomic Segment Results First 1 2 3 4 5 Next Last	Genomic Locations Advanced F Crganism C Anncaliia algerae Undeen Anncaliia algerae Undeen Anncaliia algerae Undeen	Add 20628 (a aging → Genomic Location CARR1000013.1: 1604 - 1610 (+) CARR1000013.1: 1604 - 1610 (-) CARR1000037.1: 501 - 507 (+)	Motif Motif AAACAA GTGTCGG	AGTITACA	ACAGTGGGATCC AGTGATTGGATCC GAATGGATCCCTT		Add Colu Add Colu STTCCTCCGAC GTAAACTTTGT AATTAAAAA	mns AC
Strategy: DNA Motif Genomic Segment Results Image: Comparison of the segment ID CAIR01000013 1:604-16107 CAIR01000013 1:604-16107 CAIR01000037 1:501-507.f CAIR01000037 1:501-507.f	Genomic Locations Advanced F Organism Ancaliia algerae Undeen Anncaliia algerae Undeen Anncaliia algerae Undeen Anncaliia algerae Undeen	Add 20628 (2aging CaR01000013.1: 1604 - 1610 (+) CAR01000013.1: 1604 - 1610 (+) CAR01000037.1: 501 - 507 (+) CAR01000037.1: 501 - 507 (-)	Cenomic Segme Motif AAACAA GTGTCGC TTATTATT TTTTTAAT	AGTITACA BAGGAAC TATGCAT	ACAGTGGGATCC AGTGATTGGATCC GAATGGATCCCTT VAAAGGGATCCAT		Add Colu Add Colu STTCCTCCGAC GTAAACTTTGT AATTAAAAA	mns AC TT

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

In the section 7.1 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.

My Strategies: New	Opened (1) All (1) Basket Examples Help	
(Segments) ⁽¹⁾ DNA Motif 20628 Segments Step 1	Strategy: DNA Motif* Rename Duplicate Save As Share Delete	
20628 Genomic Segments 1 Strategy: DNA Motif	Run a new Search for Genes Add Step Add contents of Basket Add existing Strategy Motify Filter by Weight ORFs Image: Content of Content	
Genomic Segment Results G First 1 2 3 4 5 Next Last Image: California segment ID Image: Califor	Advance Advance Organism Anncaliia algerae Drdeen CAIR010000 Anncaliia algerae CAIR0100000 CAI	Text (product name, notes, etc.) Gene ID(s) Organism User Comments
CAIR01000037.1:501-507:r	Add Step 2 : Organism Organism Select all clear all expand all collapse all reset to default Select all clear all expand all collapse all reset to default Select all clear all expand all collapse all reset to default Select all clear all expand all collapse all reset to default Select all clear all expand all collapse all reset to default Select all clear all expand all collapse all reset to default Select all clear all expand all collapse all reset to default Select all clear all expand all collapse all reset to default Select all clear all expand all collapse all reset to default Select all clear all expand all collapse all reset to default	Close
	Combine Genomic Segments in Step 1 with Genes in Step 2: Intersect 2	

	Add Step	8
G	nomic Colocation 🕫 🗘	
Combine Step 1 and	Step 2 using relative locations in the genome	
You had 20628 Genomic Segments in your	trategy (Step 1). Your new Genes search (Step 2) returned 35	231 Genes.
"Return each Gene from Step 2 • whose upstream region	overlaps - the exact region of a Genomic Segme	ent in Step 1 and is on either strand 🔹 "
(35231 Genes in Step)	/ \ (20628 Genomic Segments i	in Step)
Region	Region	
Gene	Genomic Segment	-
© Exact	Exact	
Opstream: 500 bp	◯ Upstream: 1000 bp	
◎ Downstream: 1000 bp	◯ Downstream: 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 G	Gene from Step 2 whose upstream region
	(12339 Genes in Step)
	Gene
	© Exact © Upstream: 500 bp
	Custom:
	begin at: start v - v 500 bp

How many genes did you get?



7.3 Using a similar sequence of steps as in part 7.2, 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.



7.4 Taking this a step further, define which of these genes do \underline{NOT} contain a BamHI 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 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 Ste GATCC	p 4								Add 2	32 Genes to E	Basket	Downl	oad 23	2 Genes	
🖂 🔻 Fi	ter result	s by specie	e s (re	sults rei	moved	by the fil	er will no	ot be con	nbined into the next	step.)							
A11	Ortholog	Encepha	alitozoo	on cunio	culi	Ence	phalitoz hellem	oon	Encophalitazoon	Enconho	litozoon	Enterocutezoon	٨	lematocio	da	Nosoma	
Results	Groups	Distinct genes E	C1 EC	2 EC3	GB - M1	Distinct genes	ATCC 50504	Swiss	intestinalis	roma	leae	bieneusi	parisii ERTm1	parisii ERTm3	sp. 1 ERTm2	ceranae	
232	106	133 3	5 32	2 32	34	21	18	15	23	2	1	12	2	1	3	0	
								1	1							۱.	
Gene	Results	Geno	me Vi	ew													
First	1234	5 Next La	st		Ad	vanced	Paging						(Add Co	lumns	D
\$	Gene II	D 🌲 Ger	nomic	Loca	tion	3	🔷 Р	roduc	t Description	3 📊				Y			
💮 EB	31_24411	ABGB	01000	099: 4	38 - 7	28 (+)	hypo	othetica	al protein								
🗇 EE	81_27581	ABGB	01000	203: 9	76 - 1	,491 (-)	hypo	othetica	al protein								
the mathematical de la construcción de la construc	31_25435	ABGB(01000	276: 1,	,036 -	1,248 (- hypo	othetica	al protein								
🗇 EE	81_26304	ABGB0 1,454)1000: (+)	351: 1,	,323 -		hypo	othetica	al protein		Sele	ect Columns					×
💮 EB	3I_26621	ABGB	01000	486: 3	58 - 5	58 (+)	hypo	othetica	al protein			l u	Jpdate	Colum	ns		
🗇 EE	81_25638	ABGB	01000	541: 2	18 - 4	30 (-)	hypo	othetica	al protein								
⊕ EE	31_25705	ABGBO	01000	850: 1	91 - 4	03 (+)	hypo	othetica	al protein			reset to	current	id all co : reset t	oliapse a to defaul	ll t	
	31_26491	ABGBU	11000	853: 3	29 - 5	41 (-)	hypo	othetica	al protein		÷ (Text, IDs, Sp	ecies				
	31_26598	ABGBU	1000	992: 5	32 - 1 75 - 0	44 (+)	hypo	othetica	al protein			Genomic Po	sition				
	01_27550	ABGBU	1001	170:4 257-5	/5-0 0-00	07 (+) 9 (+)	nypo	otnetica	ai protein			- 📃 Chromoso	ome				
	27652	ABGBO	11001	207.0 308-1	9 - 23 81 - 3	0 (+) 03 (+)	hype	antate-a	annionia ligase			- 🗹 Genomic	Locatio	n			
A FO	101 03	RU VI 201.	737· E	2 801	53 7	33 (1) 733 ()	cimi	arity to				🗏 📃 Gene Stra	and				
											÷ (🗌 Gene Attribu	tes				
											÷ (Protein Attrib	outes				
												- 🗹 Product D	escripti)	on			
												- 🔲 Molecular	Weight	t			
												🔤 Isoelectric	: Point				
												Protein Feati	ures				
												Putative Fun	ction				
												Evolution	cuon				
												Search PDB	by the r	protein s	equenc	е	
											- (GBrowse		-			
												clear all reset to	l expan	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 Tracks Preferences						
Search	1					
Landmark or Region:	Annotate R	estrictio	on Sites	- Config	ure Go	
NC_003229:162,593182,592 Search	Save Sna	pshot	Load Sna	pshot		
Data Source MicrosporidiaDB GBrowse v2.48	Scroll/Zoom	n: <mark><< </mark>	Show	/ 20 kbp 💌 🕂	<mark>> >></mark> 🖸 Flip	
Overview NC_003229 Victorial control of the second					 	
Region						
ok 10k 20k 30k 40k 50k 60k 70k 80k	90k 100k	110k	120k 130k	140k 150k 1	60k 170k 180	190k
Details						1
Calimana and a second and a second and a second	172k 173k	174	The restriction	on site plugin ger	erates a restricti	ion map on the c
★ ■ 21 Mannotated Genes (with UTRs in gray when available)	ECU02 1420 EC	102 1	This plugin v	vas written Elizal	oeth Nickerson 8	Lincoln Stein.
ECU02_1300 ECU02_1400 ECU02_1370 ECU02_1390 ECU02_140	10 ECU02 14	130	Cancel	Configure	• • • •	
	•		Select Res	triction Sites Ic	Annotate	
			Restriction	Site Display 🤇	off 🍳 on	
Select Tracks	Clear highlig	hting	🗖 Aatll	🔲 BspDI	🔲 Hpall	🗖 PspGl
	2		Acc65I	🗏 BspEl	🔲 Hpy188I	🗖 PspOMI
			🗖 Accl	🗖 BspHl	🔲 Hpy188III	🗖 Pstl
			🗖 Acll	🗖 BsrFl	🔲 Нру99І	Pvul
			🗖 Afel	BsrGl	HpyCH4III	Pvull
			🗖 Afili	BssHII	HpyCH4IV	Rsal
			🔲 Afilii	BssKl	HpyCH4V	Rsrll
			🗖 Agel	BstAPI	Kasl	Sacl
			Ahdl	BstBl	Kpnl	Sacll
			🗖 Alul	BstEll	Mbol	Sall
			AlwNI	BstNI	Mfel	Sau3Al
				BstUl	Mlul	Sau96l
			Apai			
			🗖 Apal	BstXI	Mscl	Sbfl
			Apal	BstXI	Mscl	Sbfl
			Apal	BstXI BstYI BstZ17I	Mscl Msel Msel Msll	Sbfl Scal