

Protein Motif Searches and Regular Expressions Exercise 6


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

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

a. Identify all genes annotated as hypothetical in *Phytophthora infestans*.


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](#)

- Agaricomycetes
- Chytridiomycetes
- Eurotiomycetes
- Leotiomycetes
- Oomycetes
 - Hyaloperonospora
 - Phytophthora
 - Phytophthora capsici
 - Phytophthora infestans
 - Phytophthora ramorum
 - Phytophthora sojae
 - Pythium
- Pucciniomycetes
- Saccharomycetes
- Schizosaccharomycetes
- Sordariomycetes
- Tremellomycetes
- Ustilaginomycetes
- Zygomycetes

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

Text term (use * as wildcard) 

Fields 

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

[select all](#) | [clear all](#)

Advanced Parameters

[Get Answer](#)

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.

(Genes)

Text
10874 Genes
Step 1

Add Step

Add Step

Run a new Search for
Transform by Orthology
Add contents of Basket
Add existing Strategy
Filter by assigned Weight

Genes
Genomic Segments (DNA)
Motif
SNPs
ORFs
SAGE Tags

Text, IDs, Species
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

Revise Step 2 : InterPro Domain

Organism Agaricomycetes Chytridiomycetes Eurotiomycetes Leotiomycetes Oomycetes Hyaloperonospora Phytophthora Phytophthora capsici Phytophthora infestans Phytophthora ramorum Phytophthora sojae Pythium Pucciniomycetes Saccharomycetes Schizosaccharomycetes Sordariomycetes Tremellomycetes Ustilaginomycetes Zygomycetes

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

Domain Database

Domain
Type three characters to see suggestions.
Or use * as a wildcard, like this: *-year-term

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

9 Genes from Step 2
Strategy: Text(5) Add 9 Genes to Basket | Download 9 Genes

Filter results by species (results removed by the filter will not be combined into the next step.)

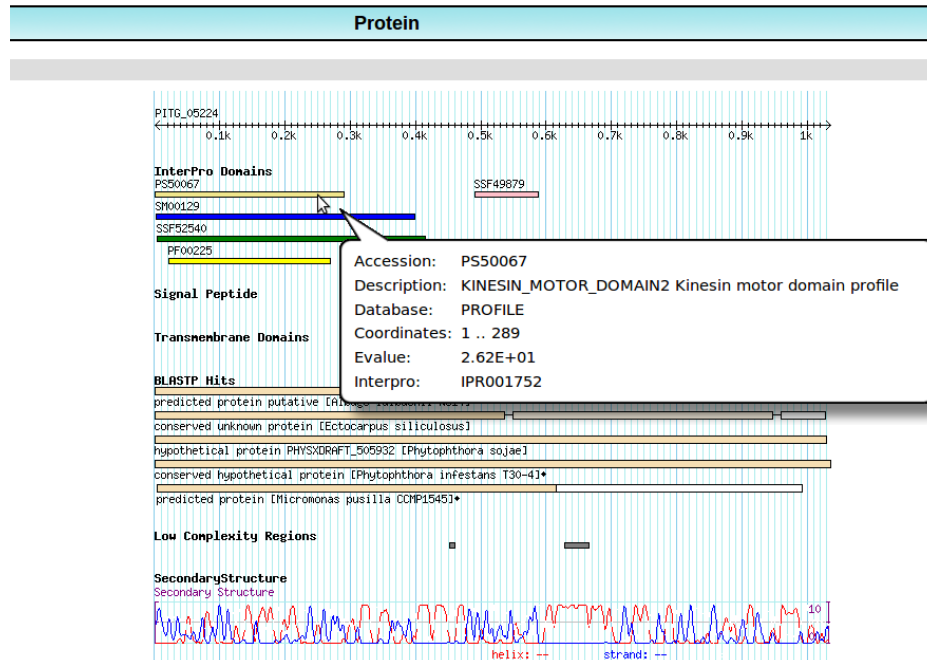
All Results	Ortholog Groups	Oomycetes					Chytridiomycetes		Zygomycetes		Agaricomycetes		Tremellomycetes					Pucciniomycetes		Ustilaginomycetes		Schizosacchar		
		Hara	Pcap	Pinf	Pram	Psoj	Pult	B.den	M.cir	R.ory	C.cin	Pchr	C.gat R265	C.gat WM276	C.neo H99	C.neo B3501	C.neo JEC21	T.mes	Pgra	M.glo	S.rei	U.may	S.jap	S.oct
9	9	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Gene Results

Gene ID	Genomic Location	Product Description	GBrowse
PITG_00478	PhyT30-4_SC0001: 2,501,858 - 2,505,298 (-)	conserved hypothetical protein	GBrowse
PITG_02538	PhyT30-4_SC0003: 2,468,713 - 2,469,267 (+)	conserved hypothetical protein	GBrowse
PITG_05224	PhyT30-4_SC0007: 1,596,587 - 1,600,159 (-)	conserved hypothetical protein	GBrowse
PITG_10158	PhyT30-4_SC0017: 2,018,080 - 2,021,529 (+)	conserved hypothetical protein	GBrowse
PITG_10474	PhyT30-4_SC0018: 1,467,527 - 1,468,650 (+)	hypothetical protein	GBrowse
PITG_11652	PhyT30-4_SC0022: 1,042,072 - 1,044,886 (+)	conserved hypothetical protein	GBrowse
PITG_16513	PhyT30-4_SC0044: 896,186 - 897,436 (+)	conserved hypothetical protein	GBrowse
PITG_19553	PhyT30-4_SC0098: 255,381 - 256,027 (+)	conserved hypothetical protein	GBrowse
PITG_21476	PhyT30-4_SC0374: 18,707 - 19,392 (+)	conserved hypothetical protein	GBrowse

Advanced Paging

- c. Go to the gene page for PITG_05224 and look at the protein feature section. Does this look like a possible motor protein?
 Hint: click on the ID for PITG_05224 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 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 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  [select all](#) | [clear all](#) | [expand all](#) | [collapse all](#) | [reset to default](#)

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

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

 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

(Genes)

