COMSATS UNIVERSITY ISLAMABAD,

SAHIWAL CAMPUS

Assignment 03 (Lab):

Topic: "Motif scan"

BIO413: Structural and Functional Bioinformatics

Submitted To:

Mam Anum

Submitted From:

Name: Iqra Masood

Registration No: SP20-BSI-008

Batch: 22

Contents

<h2 style="text-align:center">Motif Scan:</h2>

1. Introduction:

Protein motifs are most commonly extracted from an initial multiple sequence alignment, but sometimes the training sequences are not strictly homologous, or the sequences contain repeated sequences, rearrangements, or other common situations that disrupt alignment approaches.

Motif scanning means **finding all known motifs that occur in a sequence**. This form lets you paste a protein sequence, and select the collections of motifs to scan and launch the search. A document deals with the interpretation of the match scores.
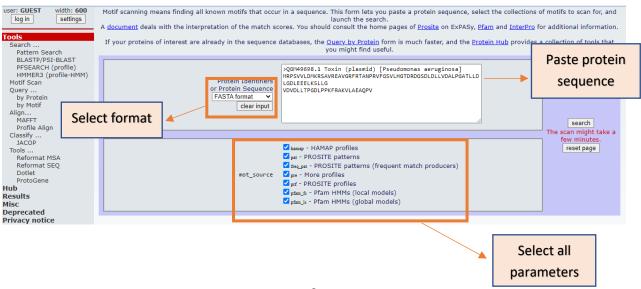
2. Methodology:

Motif Scan is a tool under MyHits. Like all MyHits tools, they operate using HitKeeper. HitKeeper was mainly written by Marco Pagni and Jörg Hau. It is a software package hit list management that contains a collection of scripts that interact with a relational database management system (RDMS).

3. Case study:

Before beginning the search, a protein sequence must be chosen. Note that Motif Scan does not scan nucleotide sequences, and only protein sequences can be used.

- Take a protein sequence of **toxin** paste it on the search bar and click on all parameters then click on the search button.
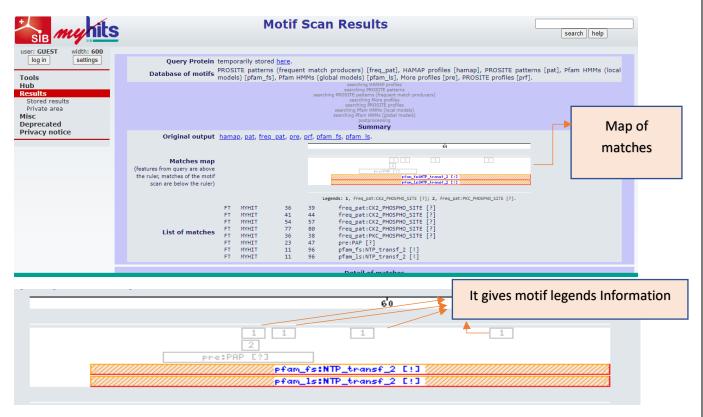- Click on examples and select fasta sequence.

3.1) Result annotation:

- It gives result from the different databases.

You should consult the home pages of Prosite on ExPASy, Pfam and InterPro for additional information.

- The names of the matches are accompanied with its status. Strong matches are denoted with a [!] while weak matches are marked with [?].

Below that, the "Detail of Matches" segment illustrates the domains and functional groups of a sequence. The position of the motif will be displayed beside it along with the status of the match. Matches that are denoted with "?" status have insufficient evidence to identify it. The tool will try to identify it anyway, but the result is questionable. **?** → tells promoter region of motif sequence is not defined and information is incomplete and have weak matches. **!** in blue color → indicates information is present but e-value is large so error chances occur.(E-value is better when it is close to 0)

3.2) Match score details:

Detail of matches

match detail	match score	motif information
TDRD	Status: **?** pos.: 36-39	Weak match
SDLD	Status: **?** pos.: 41-44	freq_pat:CK2_PHOSPHO_SITE *Casein kinase II phosphorylation site.* [entry] Legends: 1, phosphorylation.
TLLD	Status: **?** pos.: 54-57	
TPGD	Status: **?** pos.: 77-80	

RSAVREAVGRFR--TANPRVFGSVLHGTDRDG-SDLDLLVDALPG-------------------

ATLLDLGDLEEELKSLLGVDVDLLTPGDLP-------PKFRAKVLAEAQPV

Status: **!**
pos.: **11-96**
raw-score = **40.3**
N-score = **20.428**
E-value = **7.9e-14**

pfam_fs:NTP_transf_2
Nucleotidyltransferase domain
[entry]

Strong match but false positive

△ image △

Status: **!**
pos.: **11-96**
raw-score = **40.3**
N-score = **20.428**
E-value = **7.9e-14**

Pos → tells the position of the motif

N-score → normalized score

E-value → E-value provides an estimation of the number of false positives

Raw score → measures weak matches

2001 which amount to a total of 408 011 338 = 10^... residues (in 1 818 627 entries).

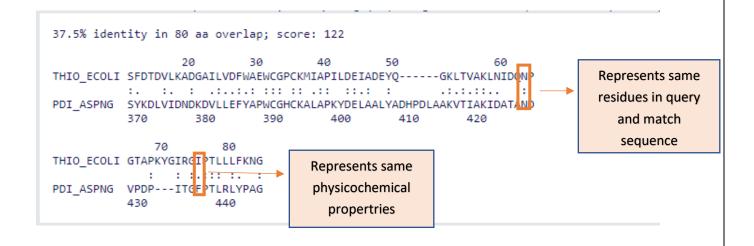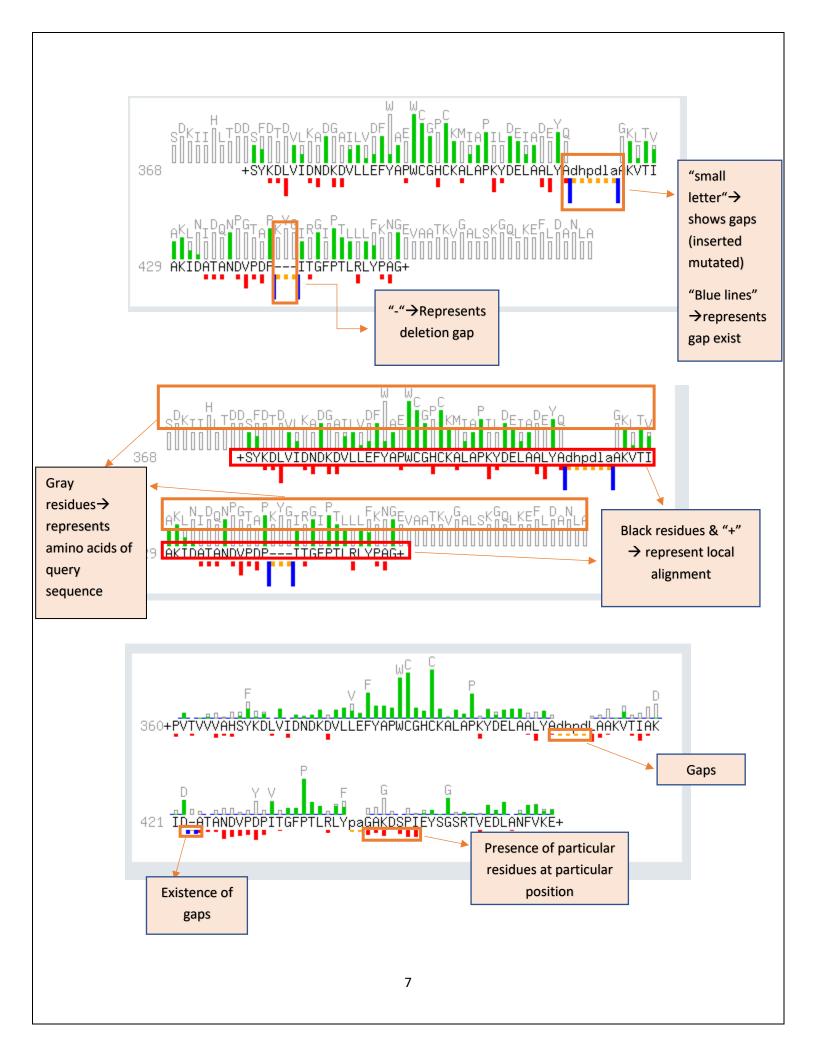| Normalized score	E-value for SwissProt version 34	E-value for the Hits database june 2001
7.0	2.1	41
8.0	0.21	4.1
9.0	0.021	0.41
10.0	0.0021	0.041

Some of these weak matches are still measured by its raw score, N-score (normalized value), and E-value despite having weak evidence. The E-value indicates how similar the motif on the sequence is to the motif obtained from one of the databases. Motifs with weak evidence tend to

5

have high E-values, indicating its dissimilarity from the motif on the sequence. The motif displayed is usually incomplete.

3.3) Match Details:

```
37.5% identity in 80 aa overlap; score: 122

              20        30        40        50              60
THIO_ECOLI SFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQ------GKLTVAKLNIDONP
           :.    :.  :  .:..:.: ::: :: .:: ::.: :       .:.:.::..  :
PDI_ASPNG  SYKDLVIDNDKDVLLEFYAPWCGHCKALAPKYDELAALYADHPDLAAKVTIAKIDATAND
           370       380       390       400       410       420

           70        80
THIO_ECOLI GTAPKYGIRGIPTLLLFKNG
           :   : :.:.:: :.  :
PDI_ASPNG  VPDP---ITGFPTLRLYPAG
           430       440
```

Represents same residues in query and match sequence

Represents same physicochemical propertries

6

368 +SYKDLVIDNDKDVLLEFYAPWCGHCKALAPKYDELAALYAdhpdlaAKVTI

429 AKIDATANDVPDF---ITGFPTLRLYPAG+

"small letter" → shows gaps (inserted mutated)

"Blue lines" → represents gap exist

"-" → Represents deletion gap

368 +SYKDLVIDNDKDVLLEFYAPWCGHCKALAPKYDELAALYAdhpdlaAKVTI

AKIDATANDVPDP---ITGFPTLRLYPAG+

Gray residues → represents amino acids of query sequence

Black residues & "+" → represent local alignment

360 +PVTVVVAHSYKDLVIDNDKDVLLEFYAPWCGHCKALAPKYDELAALYAdhpdLAAKVTIAK

421 ID-ATANDVPDPITGFPTLRLYpaGAKDSPIEYSGSRTVEDLANFVKE+

Gaps

Existence of gaps

Presence of particular residues at particular position

3.4) Graphical representation:

Matched residues

Mismatched residues

Gaps

Status: **!**
pos.: **11-96**
raw-score = **42.1**
N-score = **17.619**
E-value = **5.1e-11**

pfam_ls:NTP_transf_2
Nucleotidyltransferase domain
[entry]

RSAVREAVGRFR--TANPRVFGSVLHGTDRDG-SDLDLLVDALPG----------------------

ATLLDLGDLEEELKSLLGVDVDLLTPGDLP-------PKFRAKVLAEAQPV

4. Conclusions:

A motif scan is a tool that finds motifs on a protein sequence and attempts to identify them. It assists in the classifications of proteins by identifying these motifs or patterns on a protein sequence. The software uses HitKeeper to manage its data so that they are in sync. Not many researchers use Motifs Scan because of how niche its function is, but it is very useful in identifying functional groups which can help scientists compare unknown proteins based on these groups. The tool might see more usage if more resources are made to help users navigate it.