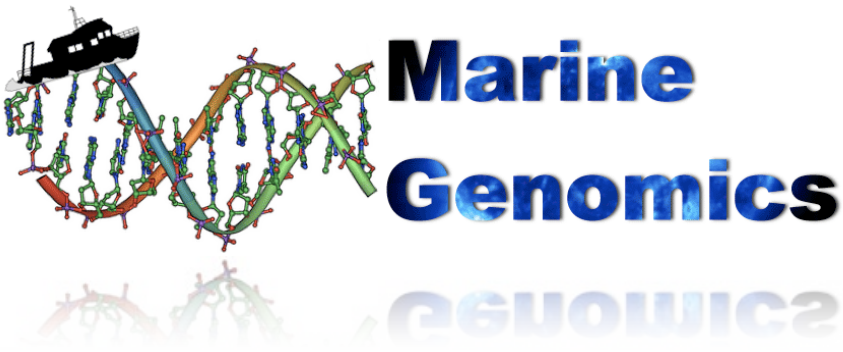


PROTOCOL

Protein Motif Analysis

compiled by John R. Finnerty

Protein Architecture: Conserved Functional Domains

- Proteins are like machines in that different parts of the protein perform different sub-functions, and together these parts allow the entire protein to perform its overall function.
- These functionally distinct parts of the protein are known as **functional domains**.
- If they are conserved across taxa, these **conserved domains** can be identified by amino acid sequence similarity.
- In the output of a BLAST search at NCBI, you will see reference to conserved domains if one or more such domains are identified.

- The top of the page provides a locus ID number, also called an “accession number,” as well as information on any publications that are associated with the sequence.

superoxide dismutase [Homo sapiens]

GenBank: AAA62278.1

FASTA Graphics

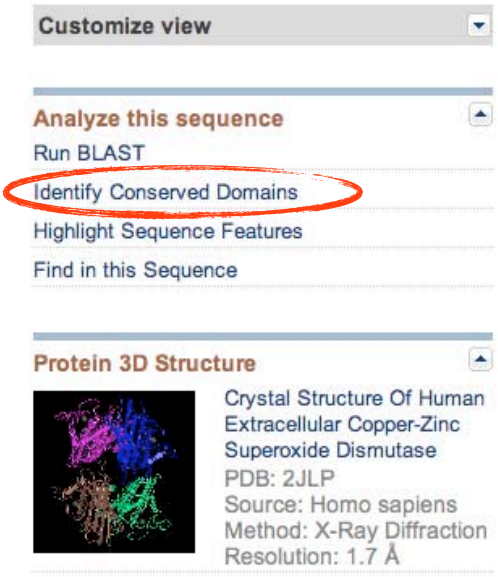
Go to:

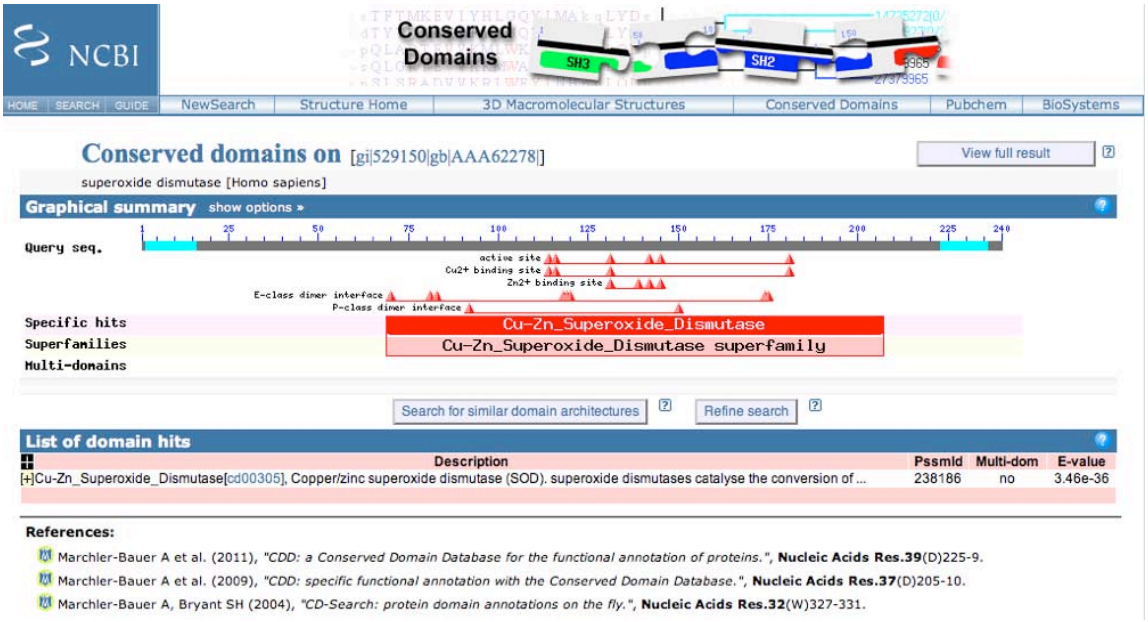
```
LOCUS       AAA62278                240 aa            linear   PRI 18-FEB-1995
DEFINITION  superoxide dismutase [Homo sapiens].
ACCESSION   AAA62278
VERSION     AAA62278.1  GI:529150
DBSOURCE    locus HSU10116 accession U10116.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 240)
  AUTHORS   Folz,R.J. and Crapo,J.D.
  TITLE     Extracellular superoxide dismutase (SOD3): tissue-specific
            expression, genomic characterization, and computer-assisted
            sequence analysis of the human EC SOD gene
  JOURNAL   Genomics 22 (1), 162-171 (1994)
```

```
     Region          69..207
                     /region_name="Cu-Zn_Superoxide_Dismutase"
                     /note="Copper/zinc superoxide dismutase (SOD). superoxide
                     dismutases catalyse the conversion of superoxide radicals
                     to molecular oxygen. Three evolutionarily distinct
                     families of SODs are known, of which the
                     copper/zinc-binding family is one. Defects in the...;
                     cd00305"
                     /db_xref="CDD:48338"
     Site            order(70,81,83,118..120,174..175)
                     /site_type="other"
                     /note="E-class dimer interface [polypeptide binding]"
                     /db_xref="CDD:48338"
     Site            order(92,150)
                     /site_type="other"
                     /note="P-class dimer interface [polypeptide binding]"
                     /db_xref="CDD:48338"
     Site            order(114,116,131,142,145,181)
                     /site_type="active"
                     /db_xref="CDD:48338"
     Site            order(114,116,131,181)
                     /site_type="other"
                     /note="Cu2+ binding site [ion binding]"
                     /db_xref="CDD:48338"
     Site            order(131,139,142,145)
                     /site_type="other"
                     /note="Zn2+ binding site [ion binding]"
                     /db_xref="CDD:48338"
     CDS             1..240
                     /gene="SOD3"
                     /coded_by="U10116.1:5085..5807"
ORIGIN
        1 mlallcscll laagasdawt gedsaepnsd saewirdmya kvteiwqevm qrrdddgtlh
       61 aacqvqpsat ldaaqprvtg vvlfrqlapr akldaffale gfptepnsss raihvhqfgd
      121 lsqgcestgp hynplavphp qhpgdfgnfa vrdgslwryr aglaaslagp hsivgravvv
      181 hageddlgrg gnqasvengn agrrlaccvv gvcgpglwer qarehserkk rrreseckaa
//
```

- The bottom of the page lists conserved regions or sites within the protein, and characterizes their known function (e.g., “polypeptide binding”, “ion binding”).
- The full protein sequence is also given.

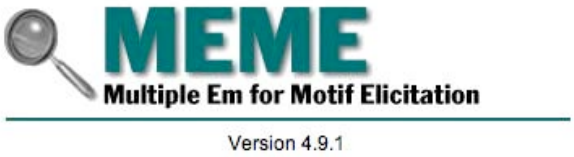
- At the top of the page, on the far right, below the link to **Run BLAST**, you can click on the link for **Identify Conserved Domains**.





How do I find conserved protein domains if not all of my sequences are annotated?

- Computer programs can detect conserved regions of proteins (known as motifs) based solely on their amino acid sequences.
- The strong conservation of a motif over evolutionary time suggests (1) that it may have an important function and (2) that its sequence is therefore be constrained by stabilizing selection.
- One popular program for identification of such conserved motifs is **MEME. Multiple Em for Motif Elicitation**



Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California, 1994.

Assemble a file of **amino acid sequences** in **FASTA format**

- Don't include only sequences from your focal taxon (e.g., *Astrangia*).
- Also include annotated sequences from well-studied model systems (e.g., *Nematostella*, *Acropora*, *Drosophila*, and/or vertebrates).
- Paste your amino acid sequences into a text file using the FASTA format.

>Sequence1Name[return]
MAGITRVAFFEDRWSACV.....[return]
>Sequence2Name [return]
MAGLTRVAYFEDRWTACV..... [return]
>Sequence2Name[return]
MLGITRVAFFDDRWTACV..... [return]

Obtaining **amino acid sequences** for an annotated protein sequence on NCBI

superoxide dismutase [Homo sapiens]

Click on the FASTA link.

GenBank: AAA62278.1

FASTA Graphics

Go to:

```
LOCUS       AAA62278                240 aa            linear   PRI 18-FEB-1995
DEFINITION  superoxide dismutase [Homo sapiens].
ACCESSION   AAA62278
VERSION     AAA62278.1  GI:529150
DBSOURCE    locus HSU10116 accession U10116.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 240)
  AUTHORS   Folz,R.J. and Crapo,J.D.
  TITLE     Extracellular superoxide dismutase (SOD3): tissue-specific
            expression, genomic characterization, and computer-assisted
            sequence analysis of the human EC SOD gene
  JOURNAL   Genomics 22 (1), 162-171 (1994)
```

Obtaining **amino acid sequences** for an annotated protein sequence on NCBI

Protein | Protein | Advanced

Display Settings: FASTA

superoxide dismutase [Homo sapiens]

GenBank: AAA62278.1

GenPept Graphics

```
>gi|529150|gb|AAA62278.1| superoxide dismutase [Homo sapiens]
MLALLCSCLLLAAGASDAWTGEDSAEPNSDSAEWIRDMYAKVTEIWQEVMQRRDDDGTLHAACQVQPSAT
LDAAQPRVTGVVLFRQLAPRAKLDAFFALEGFPTEPNSSSRAIHVHQFGDLSQGCESTGPHYNPLAVPHP
QHPGDFGNFAVRDGSLWRYRAGLAASLAGPHSIVGRAVVVHAGEDDLGRGGNQASVENGNAGRRLACCVV
GVCGPGLWERQAREHSERKKRRRESECKAA
```

Obtaining **amino acid sequences** for an an unannotated sequence (e.g., from *Astrangia*)

- Find sequence by BLASTing vs. *Astrangia* at skinnybastard.bu.edu (choose StellaBase).

Obtaining **amino acid sequences** for an an unannotated sequence (e.g., from *Astrangia*)

- Paste Query sequence in text box.
- Select "tblastn" as program
- Select "Astrangia V2" as database

Obtaining **amino acid sequences** for an an unannotated sequence (e.g., from *Astrangia*)

- Paste Query sequence in text box.
- Select "tblastn" as program
- Select "Astrangia V2" as database


```
Database: Velvet Oases assembly using a multikmer merge approach of Astrangia
poculata.
           763,648 sequences; 1,098,732,827 total letters


Query=  GI|529150|GB|AAA62278.1|  SUPEROXIDE DISMUTASE [HOMO SAPIENS]
Length=240

                                                                    Score     E
Sequences producing significant alignments:                        (Bits)  Value

  Locus_3843_Transcript_6/44_Confidence_0.194_Length_861             101     1e-20
  Locus_3843_Transcript_5/44_Confidence_0.208_Length_855             101     1e-20
  Locus_3843_Transcript_30/44_Confidence_0.208_Length_849            100     2e-20
  Locus_3843_Transcript_11/44_Confidence_0.264_Length_1023           99.8    3e-20
  Locus_3843_Transcript_9/44_Confidence_0.194_Length_855             99.8    3e-20
  Locus_3843_Transcript_28/44_Confidence_0.208_Length_855            99.8    4e-20
  Locus_3843_Transcript_35/44_Confidence_0.181_Length_855            99.4    4e-20
  Locus_3843_Transcript_33/44_Confidence_0.194_Length_861            99.4    4e-20
  Locus_3843_Transcript_12/44_Confidence_0.306_Length_1023           99.4    4e-20
  Locus_3843_Transcript_8/44_Confidence_0.236_Length_855             99.4    4e-20

☑> Locus_3843_Transcript_6/44_Confidence_0.194_Length_861
Length=861

 Score =  101 bits (252),  Expect = 1e-20, Method: Compositional matrix adjust.
 Identities = 58/143 (40%), Positives = 76/143 (53%), Gaps = 11/143 (7%)
 Frame = +2

Query  77   RVTGVVLFRQLAPRAKLDAFFALEGFPTEPNSSSRAIHVHQFGDLSQGCESTGPHYNPLA  136
            ++ GV+ F Q A   +      + G  T         HVHQFGD + GC S GPH+NP
Sbjct  89   KLMGVIHFEQEAEGKEC----KITGEVTGLTEGKHGFHVHQFGDGTNGCTSAGPHFNPTG  256
```

- Note the correct reading frame.
- Click the checkbox of the sequence you want to download.

```
Database: Velvet Oases assembly using a multikmer merge approach of Astrangia
poculata.
           763,648 sequences; 1,098,732,827 total letters


Query=  GI|529150|GB|AAA62278.1|  SUPEROXIDE DISMUTASE [HOMO SAPIENS]
Length=240
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

  Locus_3843_Transcript_6/44_Confidence_0.194_Length_861              101     1e-20
  Locus_3843_Transcript_5/44_Confidence_0.208_Length_855              101     1e-20
                                                                      100     2e-20
                                                                      99.8    3e-20
                                                                      99.8    3e-20
                                                                      99.8    4e-20
                                                                      99.4    4e-20
                                                                      99.4    4e-20
                                                                      99.4    4e-20
                                                                      99.4    4e-20
```

- Note the correct reading frame.
- Click the checkbox of the sequence you want to download.
- At the bottom of the page, click "Selected Fastas."

```
>Locus_3843_Transcript_6/44_Confidence_0.194_Length_861

Length=861

 Score =  101 bits (252),  Expect = 1e-20, Method: Compositional matrix adjust.
 Identities = 58/143 (40%), Positives = 76/143 (53%), Gaps = 11/143 (7%)
 Frame = +2

Query  77   RVTGVVLFRQLAPRAKLDAFFALEGFPTEPNSSSRAIHVHQFGDLSQGCESTGPHYNPLA  136
            ++ GV+ F Q A   +      + G  T          HVHQFGD + GC S GPH+NP
Sbjct  89   KLMGVIHFEQEAEGKEC----KITGEVTGLTEGKHGFHVHQFGDGTNGCTSAGPHFNPTG  256
```

```
>Locus_3843_Transcript_6/44_Confidence_0.194_Length_861
TCAAAATCTTGTATCCATGGTGAGTTGTTTATTTATTTTTCATTTTTAGGTTGTACCAGTGCTGGTCCCCACTTTAATCC
TACTGGTAAAACTCATGGGTGTGATCCACTTTGAACAGGAGGCCGAAGGAAAAGAGTGTAAGATTACTGGGGAAGTAACA
GGTCTTACTGAAGGAAAACATGGATTTCATGTCCATCAGTTTGGTGATGGCACAAATGGTTGTACCAGTGCTGGTCCCCA
CTTTAATCCTACTGGTAAAACTCATGGAGGTCCAGATGATGAAATACGTCATTATGGGGACCTTGGTAACATCACAGCAG
ATAAAGATGGTAAAGCAAAAATTGACATGACGGACAAACTAGTTTCCATTATTGGAAAGGACTCTGTTGTTGGACGCACA
ATTGTGGTACATGCCAAGGTAGATGACTTAGGAAAGGGTGGTGATCAGGAGAGTCTGAAGACTGGCAATGCTGGTGCACG
CTGGGCCTGTGGAGTGATTGGCATTACCAAGTAAACAGCACCCCTGGCCAAGTCGTGGTGTGCATTATTTTGACGATTGG
AGAGACTATATGTGGCTGTAATGCCTTCAGACTTAAAGTCTGCATTAGTAACTAACAAGAGACATGTACTGAGTAACCAA
TAAAGTTCAACTTAATGTTTTCAGTTGTTCTGCGGTTTGTTGATTAGAAGTGCGATGAAGACTCAGGTTGACTGGCAAAT
GAGCCACAACTTGAACCTGCAAAGATATTTTCTATCTAGCCGCTTTGGTTTCTCTGCAAAAGTACCTGAAAAAAGGTTCC
ATTGAGTTAACACTTTTACACAAAATTTAATTTCATGTTAATTATCTTGCAGTGCAAAACT
```

- Use an online translation tool (e.g., www.expasy.org) to obtain the predicted amino acid sequence.

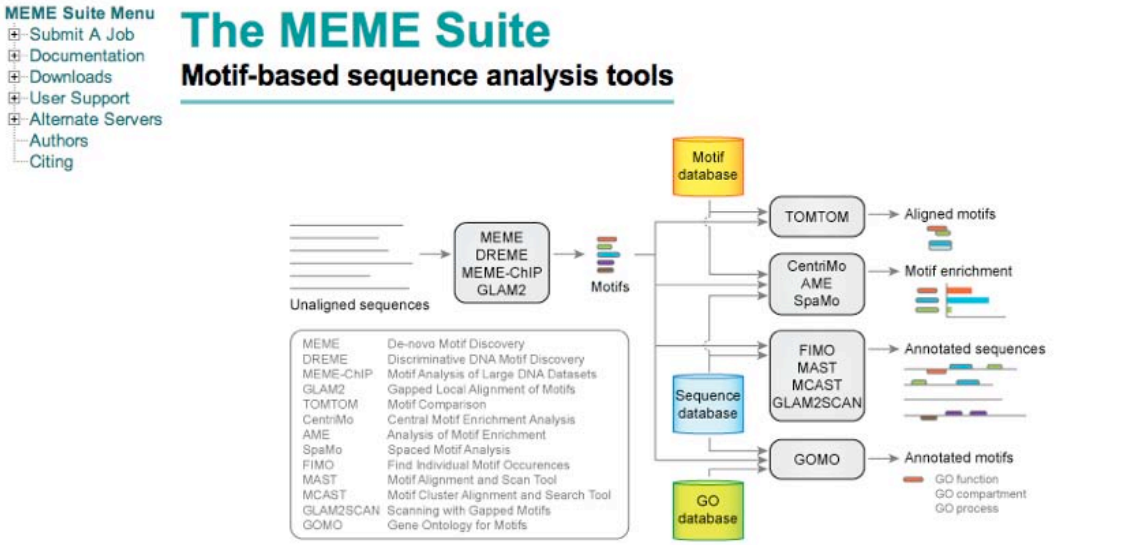
ExPASy Bioinformatics Resource Portal — Translate — Home | Contact

Translate Tool

Open reading frames are highlighted in red. Please select one of the "*M*ethionine" or one of the highlighted residues following a **Stop** codon (or the beginning of the sequence).

This will create a virtual Swiss-Prot entry, comprising the residues from your chosen start position up to the following **Stop** codon.

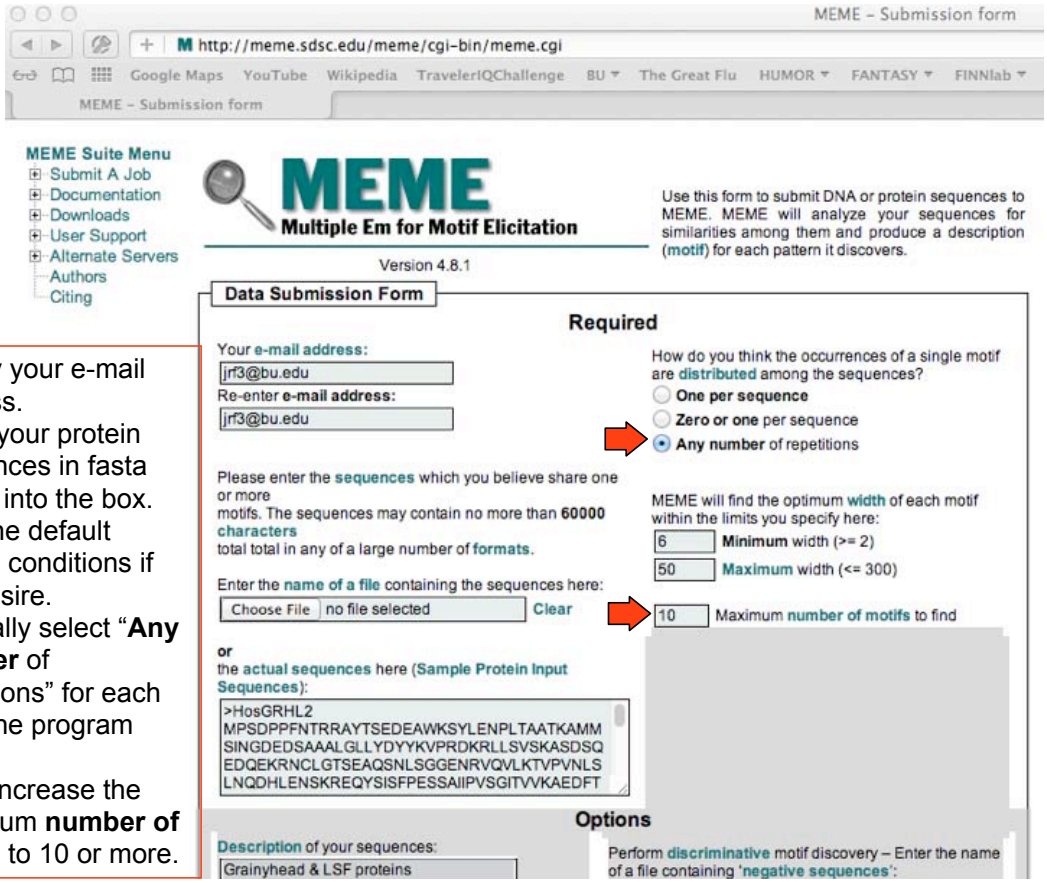
X X X X X X X X X X Q N L V S M V S C L F I F H F Stop V V P V L V P T L I L L V K L M G V I H F E Q E A E
G K E C K I T G E V T G L T E G K H G F H V H Q F G D G T N G C T S A G P H F N P T G K T H G G P D D
E I R H Y G D L G N I T A D K D G K A K I D M T D K L V S I I G K D S V V G R T I V V H A K V D D L G K G G
D Q E S L K T G N A G A R W A C G V I G I T K Stop T A P L A K S W C A L F Stop R L E R L Y V A V M P S
D L K S A L V T N K R H V L S N Q Stop S S T Stop C F Q L F C G L L I R S A M K T Q V D W Q M S H N L N
L Q R Y F L S S R F G F S A K V P E K R F H Stop V N T F T Q N L I S C Stop L S C S A K



The MEME Suite allows you to:

- discover motifs using MEME, DREME (DNA only) or GLAM2 on groups of related DNA or protein sequences,
- search sequence databases with motifs using MAST, FIMO, MCAST or GLAM2SCAN,
- compare a motif to all motifs in a database of motifs,
- associate motifs with Gene Ontology terms via their putative target genes, and
- analyse motif enrichment using SpaMo or CentriMo.

http://meme.nbcr.net/meme/



- Supply your e-mail address.
- Paste your protein sequences in fasta format into the box.
- Alter the default search conditions if you desire.
- I typically select "**Any number** of repetitions" for each motif the program finds
- I also increase the Maximum **number of motifs** to 10 or more.

Your job id is: **app1347761391297**
You can view your job results at: http://meme.nbcr.net/meme4_8_1/cgi-bin/querystatus.cgi?jobid=app134776139
You can view server activity here.

Description

```
Grainyhead & LSF proteins
```

Settings

Sequence file	sequences
Distribution of motif occurrences	Any number of repetitions
Number of different motifs	10
Minimum motif width	6
Maximum motif width	50

Sequences

Type of Sequences	protein
Count of Sequences	32
Shortest Sequence (residues)	50
Longest Sequence (residues)	1064
Average Length (residues)	540.0
Total Length (residues)	17279

You will also receive a confirming message at your email address: **jrf3@bu.edu**.

- After you select "Start Search", this summary window will appear.
- It summarizes
  1. the **Description** you provided,
  2. the **Settings** you specified, and
  3. a description of the **Sequences** you submitted
- Click the link to view your **job results**.

MEME Job - Done

You may bookmark this page and return to it later.

- When the job is done, this window will appear.
- Click on MEME html output to see the motifs that were identified.

Results

- MEME html output
- MEME xml output
- MEME txt output
- MAST html output
- MAST xml output
- MAST txt output
- Input sequences

Status Messages

- Parsing arguments
- Arguments ok
- Starting meme

```
meme sequences -protein -oc . -nostatus -time 7200 -maxsize 60000 -mod anr -nmo
```

- meme ran successfully in 354.45 seconds
- Starting mast

```
mast meme.xml sequences -oc . -nostatus
```

- mast ran successfully in 0.29 seconds

Discovered Motifs

Motif Overview

Motif 1

- 2.7e-407
- 28 sites

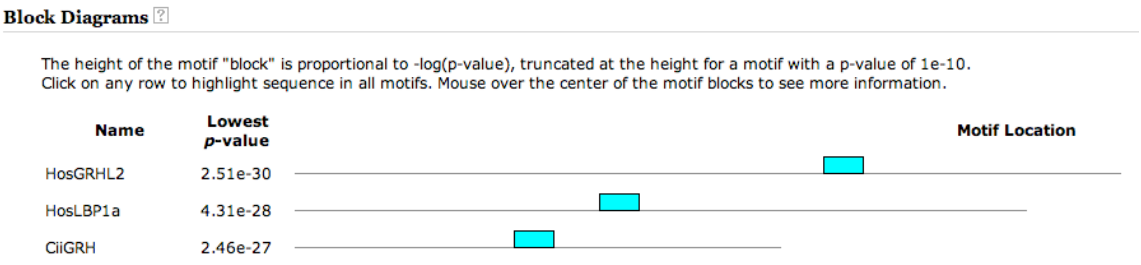
- For each **Discovered Motif**, the overview will provide
  - a **E-value**, here 2.7e-407, which "is an estimate of the expected number of motifs with the given log likelihood ratio (or higher), and with the same width and site count, that one would find in a similarly sized set of random sequences. (In random sequences each position is independent with letters chosen according to the background letter frequencies.)"
  - a count of the **number of sites observed** (here 28, which amounts to 1 site per protein sequence provided)
  - a **sequence LOGO** representing the conserved motif
    - each position in the LOGO diagram shows the amino acids that are observed to occur in that position.
    - the height of the letter is proportional to how many times that amino acid was observed in that position. The LOGO diagram above reflects the fact that cysteine (C) was always observed at position 5 of Motif 1.
- Click on the link to Motif 1 and you will see an amino acid alignment of the motif from all the protein sequences in which it was identified

Click on any row to highlight sequence in all motifs.

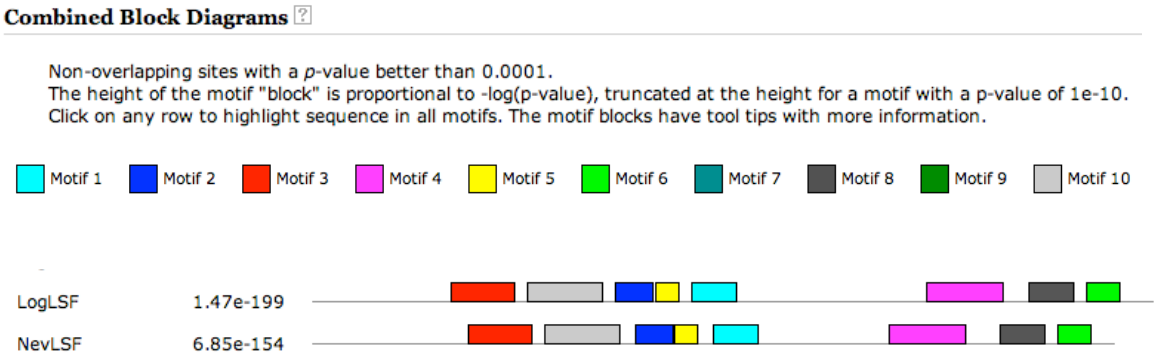
Name	Start	*p*-value	Sites
EdGRH1	235	2.13e-30	DVDNPDAEPV RAFCQIKVFRDKGAERKNKDESRSAERR MQKWMKQNPI
HosGRHL2	391	2.51e-30	SYNNRSNKPI RAYCQIKVFCDKGAERKIRDEERKQNRK KGKGQASQTQ
NevGRH1	169	5.53e-30	DVDNPDAEPV RAFCQIKVFRDKGAERKNKDESKSAERR
DrmGRH	791	7.52e-30	FEDPRDTAVF RGYCQIKVFCDKGAERKTRDEERRAAKR KMTATGRKKL
BrfLSF	215	1.97e-28	DSYSIEEEHL RSASCQIKVFKPKGADRKIKTDREKMEKK PDKDKYQPSY
CapGRH	131	3.33e-28	FEDTSSATPI RGYCQVKVFCDKGAERKTRDEERRKDKS KPDGMSLTAA
HosLBP1a	226	4.31e-28	NENGEYTDHL RSASCQIKVFKPKGADRKQKTDREKMEKR TAHEKEKYQP
LogLSF	250	4.90e-28	SHEEEEEKLL RSASCQVKVFKPKGADRKHKTDREKMDKR SESEKEKYQP
AmqGRH	429	6.32e-28	DRMTEHDEPS RAYCRVKIFRDKGAERKNKDETKSVERR LQKFIRSYNS
PhbLSFL1	248	6.32e-28	GTQVVGEQHI EKAYCRIKLFRDKGAERKNKDDAKHIERK NGEPHPLWLT
CapLSF	248	3.96e-27	IGDIHPARIV RCSSCQVKVFKPKGADRKHKTDRERIEKR SETEKLFFRP
DapLSF	263	1.26e-26	GDGDGTPKRL RVAGCQIKVFKLKGADRKHKQDREKIYKR PMVEQEKYQP
CilLSF	218	3.85e-26	QNNNEYGRYI RSASCQIKVFKPKGADRKQKTDKDKMERR TAQEKLKYQP
PhbLSFL2	540	6.63e-26	DSQYGTYDYV ESCFCKIKLFRDKGAERKIKDDAKQINRH LEKLFSEGNH

- The motif is the same length in each protein (28 amino acids).
- The amino acid sequence is not strictly conserved, but often the same class of amino acid (e.g., polar, non-polar, positively charged, or negatively charged) is found at the same position in the motif in most or all of the sequences (indicated by the color). For example, position 16 is always a negatively charged amino acid (aspartic acid [D] or glutamic acid [E])

- The Block Diagram for each motif depicts the motif as a colored block which has the following properties
  - its **height is inversely proportional to its p-value**, &
  - its **location is shown relative to the rest of the protein** (the entire protein is represented by the fine line)



- At the bottom of the page, examine the **Combined Block Diagrams** to see how the overall protein architecture (the relative location of conserved motifs) compares across proteins.



- Notice that the LSF proteins of the snail *Lottia* (LogLSF) and the anemone *Nematostella* (NevLSF) differ in length, but they share the same 8 conserved motifs in the same relative order.
- This conservation of protein architecture suggests that they two proteins are performing the same functions.