

Tutorial 7.

Motifs and Domains

The Chinese University of Hong Kong

BMEG3102 Bioinformatics

TA: Yizhen Chen



Agenda

- Introduction to motifs and domains
 - Definition
 - Representation
 - Exact
 - Probabilistic
- Motif demonstration
 - Databases
 - Web tools

Definition of Motifs and Domains

1. Occur frequently

- May not be exactly the same in different occurrences.
- Are unlikely to occur 'by chance'; they are 'over-represented'.

2. Usually have known or predicted functional roles

3. Are evolutionarily conserved

- Examples:

- DNA motif, protein domain

Transcription Factor

- Transcription factors are proteins that contain one or more DNA-binding domains, which attach to **specific sequences of DNA** adjacent to the genes that **they regulate**. (from wiki)
- Representations:
 - Consensus sequence representation
 - Degenerate sequence in IUPAC code
 - Regular expression representation
 - Position weight matrix (PWM)
 - Sequence logo

How to represent a motif?

- Given the occurrences of a motif below. Use consensus sequence, degenerate sequence, regular expression, PWM, and sequence logo to represent the motif.

HEM13	CCCATTGTTCTC
HEM13	TTTCTGGTTCTC
HEM13	TCAATTGTTTAG
ANB1	CTCATTGTTGTC
ANB1	TCCATTGTTCTC
ANB1	CCTATTGTTCTC
ANB1	TCCATTGTTCGT
ROX1	CCAATTGTTTGT

Example: Consensus Sequence

- Consensus sequences from the majority vote.

CCCATTTGTTCTC
TTCTGGTTCTC
TCAATTGTTTAG
CTCATTTGTTGTC
TCCATTTGTTCTC
CCTAATTGTTCTC
TCCATTTGTTCTGT
CCAATTGTTTGTG

Consensus sequence: CCCATTTGTTCTC
or TCCATTTGTTCTC

Example: Degenerate Sequence

- Degenerate sequence in IUPAC code

```
CCCATTTGTTCTC
TTTCTGGTTCTC
TCAATTGTTTAG
CTCATTTGTTGTC
TCCATTTGTTCTC
CCTATTTGTTCTC
TCCATTTGTTTCGT
CCAATTTGTTTTG
```

Degenerate sequence:

```
YYHMTKGTTBDB
```

Code	Description
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
R	Purine (A or G)
Y	Pyrimidine (C or T)
M	A or C
K	G or T
W	A or T
S	C or G
B	C, G or T (not A)
D	A, G or T (not C)
H	A, C or T (not G)
V	A, C or G (not T)
N	Any base (A, C, G or T)

Define a Regular Expression

- A regular expression is a specific pattern that provides concise and flexible means to match text strings.
 - For example, we can define:

Metacharacter	Description
.	Matches any single character. e.g.: A.C matches "AAC", "ACC", "AGC" or "ATC".
[]	A bracket expression. Matches a single character that is contained within the brackets. e.g.: A[ACT] matches "AA", "AC", or "AT".
A{m, n} or (A){m, n}	Repeat a substring, i.e.: A, at least m and at most n times. e.g.: A{2,3} matches "AA" or "AAA". (GT){3} matches "GTGTGT" only.

Example: Regular Expression

CCCATTGTTCTC
TTTCTGGTTCTC
TCAATTGTTTAG
CTCATTGTTGTC
TCCATTGTTCTC
CCTATTGTTCTC
TCCATTGTTCGT
CCAATTGTTTGT

One possible representation in regular expression:

`[CT]{2}[ACT][AC]T[GT]GT{2}[CGT][AGT][CGT]`

Example: Position Weight Matrix

- Position Weight Matrix is a matrix that shows a weighted match to a list of sequences.

Nucleotide Frequency

Number of Occurrence

A	0	0	2	7	0	0	0	0	0	0	1	0
C	4	6	4	1	0	0	0	0	0	5	0	5
G	0	0	0	0	0	1	8	0	0	1	1	2
T	4	2	2	0	8	7	0	8	8	2	6	1

Position Weight Matrix

Probability of Occurrence

A	0	0	0.25	0.88	0	0	0	0	0	0	0.13	0
C	0.5	0.75	0.5	0.13	0	0	0	0	0	0.63	0	0.63
G	0	0	0	0	0	0.13	1	0	0	0.13	0.13	0.25
T	0.5	0.25	0.25	0	1	0.88	0	1	1	0.25	0.75	0.13

Pseudo-count in PWM

- About pseudo-count:
 - Pseudo-count is a small number added to each count on the occurrence of nucleotides.
 - Adding pseudo-count is to alleviate problems due to the small sample size

A	0	0	2	...
C	4	6	4	...
G	0	0	0	...
T	4	2	2	...

A	0	0	0.25	...
C	0.5	0.75	0.5	...
G	0	0	0	...
T	0.5	0.25	0.25	...



A	1	1	3	...
C	5	7	5	...
G	1	1	1	...
T	5	3	3	...

A	0.08	0.08	0.25	...
C	0.42	0.58	0.42	...
G	0.08	0.08	0.8	...
T	0.42	0.25	0.25	...

Example: Position Weight Matrix

- Based on the PWM, what is the probability that we have a motif YYCATTGTTCTC (in IUPAC)?

A	0	0	0.25	0.88	0	0	0	0	0	0	0.13	0
C	0.5	0.75	0.5	0.13	0	0	0	0	0	0.63	0	0.63
G	0	0	0	0	0	0.13	1	0	0	0.13	0.13	0.25
T	0.5	0.25	0.25	0	1	0.88	0	1	1	0.25	0.75	0.13

- Answer:

$$\begin{aligned}
 & (0.5+0.5)(0.75+0.25)(0.5)(0.88)(1) \\
 & (0.88)(1)(1)(1) (0.63)(0.75)(0.63) \\
 & = 0.11525976
 \end{aligned}$$

Code	Description
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
R	Purine (A or G)
Y	Pyrimidine (C or T)
M	A or C
K	G or T
W	A or T
S	C or G
B	C, G or T (not A)
D	A, G or T (not C)
H	A, C or T (not G)
V	A, C or G (not T)
N	Any base (A, C, G or T)

Example: Sequence Logo

- Sequence logo is the graphical representation of the sequence conservation
- Height of nucleotide j at position $i = p_{j,i} \times h_i$

$$h_i = 2 + \sum_{x \in \{A,C,G,T\}} p_{i,x} \log_2 p_{i,x} - \frac{4 - 1}{2n \ln 2}$$

n : number of sequences

$0 \times \log_2 0$ is defined as 0

$$\begin{aligned} H_7 &= 2 + 0 \log_2 0 + 0 \log_2 0 \\ &\quad + 1 \log_2 1 + 0 \log_2 0 \\ &\quad - 3/(2 \times 8 \times \ln(2)) \\ &= 1.7295 \end{aligned}$$

$$\text{Height of A} = 0 \times 1.7295 = 0$$

$$\text{Height of C} = 0 \times 1.7295 = 0$$

$$\text{Height of G} = 1 \times 1.7295 = 1.7295$$

$$\text{Height of T} = 0 \times 1.7295 = 0$$



A	0	0	0.25	0.88	0	0	0	0	0	0	0.13	0
C	0.5	0.75	0.5	0.13	0	0	0	0	0	0.63	0	0.63
G	0	0	0	0	0	0.13	1	0	0	0.13	0.13	0.25
T	0.5	0.25	0.25	0	1	0.88	0	1	1	0.25	0.75	0.13

Example: Sequence Logo

- Height of nucleotide j at position $i = p_{j,i} \times h_i$

- $$h_i = 2 + \sum_{x \in \{A,C,G,T\}} p_{i,x} \log_2 p_{i,x} - \frac{4 - 1}{2n \ln 2}$$

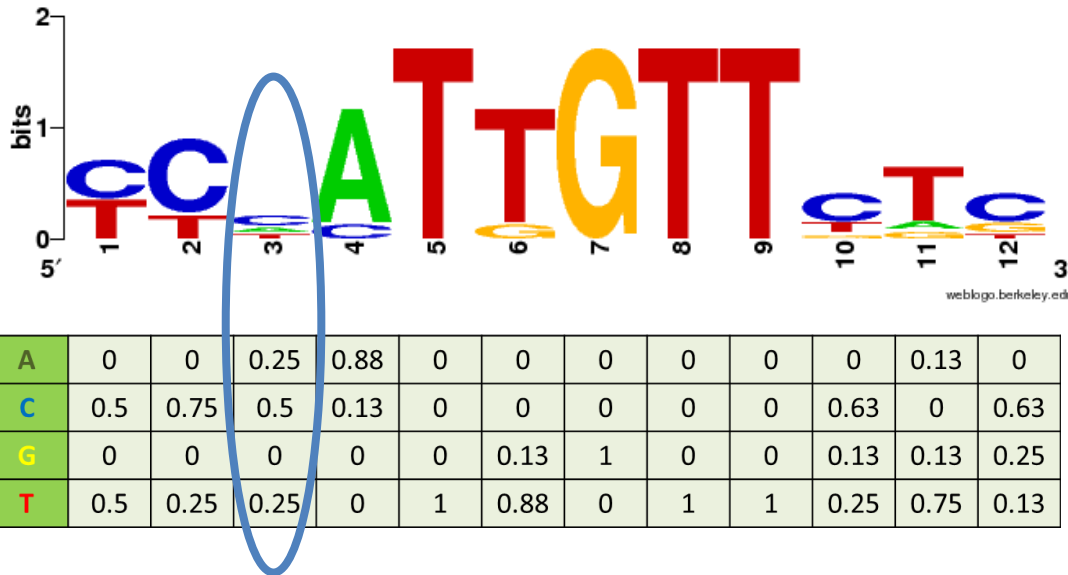
$$\begin{aligned} H_3 &= 2 + 0.25 \log_2 0.25 \\ &\quad + 0.5 \log_2 0.5 \\ &\quad + 0 \log_2 0 \\ &\quad + 0.25 \log_2 0.25 \\ &\quad - 3/(2 \times 8 \times \ln(2)) \\ &= 0.2295 \end{aligned}$$

$$\begin{aligned} \text{Height of A} &= 0.25 \times 0.2295 \\ &= 0.0574 \end{aligned}$$

$$\begin{aligned} \text{Height of C} &= 0.5 \times 0.2295 \\ &= 0.1147 \end{aligned}$$

$$\text{Height of G} = 0 \times 0.2295 = 0$$

$$\begin{aligned} \text{Height of T} &= 0.25 \times 0.2295 \\ &= 0.0574 \end{aligned}$$



Properties of the Sequence Logo

- The height of the character indicates the probability of having such nucleotide in the motif.
- Nucleotide with the highest probability was drawn on the top at each site.
- The total height of a site shows the randomness in seeing different nucleotides. Larger height means the site is more conserved (i.e. less randomness).

Exercise: Motif representation

- Given the following sequence alignment:

CCTGGC
 CCTGGA
 CCTGGC
 CCTGGC
 CCGGGC
 CTTGGC
 CTTGGC
 CCGGGC
 CTTGGC
 CTTGGC

Code	Description
A	Adenine
C	Cytosine

Metacharacter	Description
.	Matches any single character. e.g.: A.C matches "AAC", "ACC", "AGC" or "ATC".
[]	A bracket expression. Matches a single character that is contained within the brackets. e.g.: A[ACT] matches "AA", "AC", or "AT".
A{m, n} or (A){m, n}	Repeat a substring, i.e.: A, at least m and at most n times. e.g.: A{2,3} matches "AA" or "AAA". (GT){3} matches "GTGTGT" only.

B	C, G or T (not A)
D	A, G or T (not C)
H	A, C or T (not G)
V	A, C or G (not T)
N	Any base (A, C, G or T)

- Write down the consensus sequence representation, degenerate sequence, regular expression, PWM.

Answer: Motif representation

- Consensus sequence: CCTGGC
- Degenerated sequence in IUPAC code: CYKGGM
- One possible regular expression:
C[CT][GT]GG[AC]
- Position weight matrix:

A	0	0	0	0	0	0.1
C	1	0.6	0	0	0	0.9
G	0	0	0.2	1	1	0
T	0	0.4	0.8	0	0	0

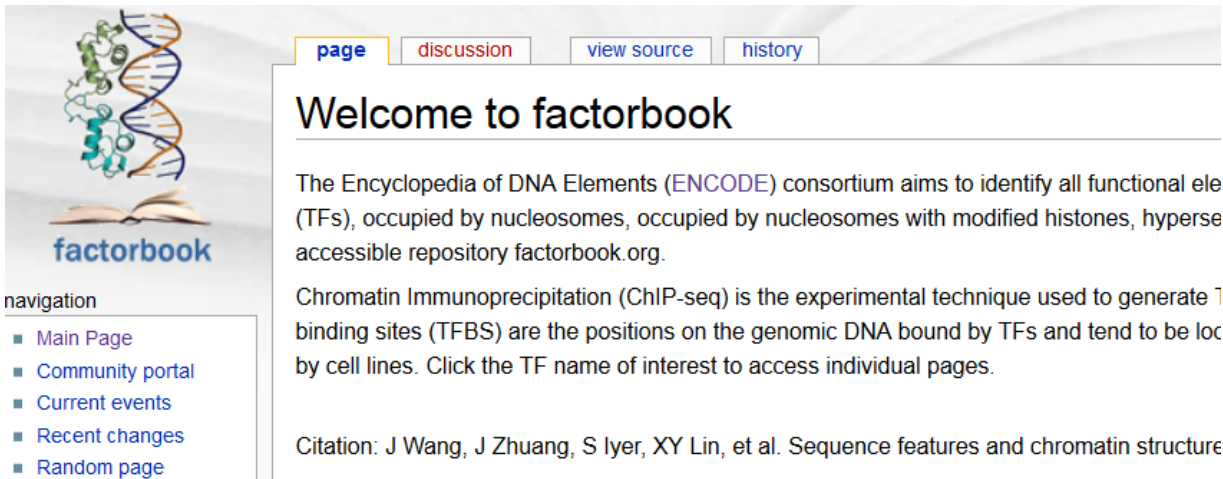
Databases and web tools

- TFBS databases:
 - Factorbook
 - <http://www.factorbook.org/human/chipseq/tf/>
 - JASPAR
 - <http://jaspar.genereg.net/>
 - Transfac
 - <http://genexplain.com/transfac/>
 - Hocomoco
 - <http://hocomoco11.autosome.ru/>
- Motif-based sequence analysis tools
 - The MEME Suite
 - <http://meme-suite.org/>
 - Motif Discovery, Motif Enrichment, Motif Scanning, Motif Comparison,

Demonstration on Databases

- FactorBook:

<http://www.factorbook.org/mediawiki/>



The screenshot shows the FactorBook website. On the left is a logo featuring a DNA double helix and the text 'factorbook'. Below the logo is a 'navigation' menu with links: Main Page, Community portal, Current events, Recent changes, and Random page. The main content area has tabs for 'page', 'discussion', 'view source', and 'history'. The title is 'Welcome to factorbook'. The text describes the Encyclopedia of DNA Elements (ENCODE) consortium and its goal to identify functional elements (TFs) in the genome. It mentions that Chromatin Immunoprecipitation (ChIP-seq) is used to generate binding sites (TFBS) and that users can click on TF names to access individual pages. A citation is provided at the bottom: 'Citation: J Wang, J Zhuang, S Iyer, XY Lin, et al. Sequence features and chromatin structure'.

- JASPAR

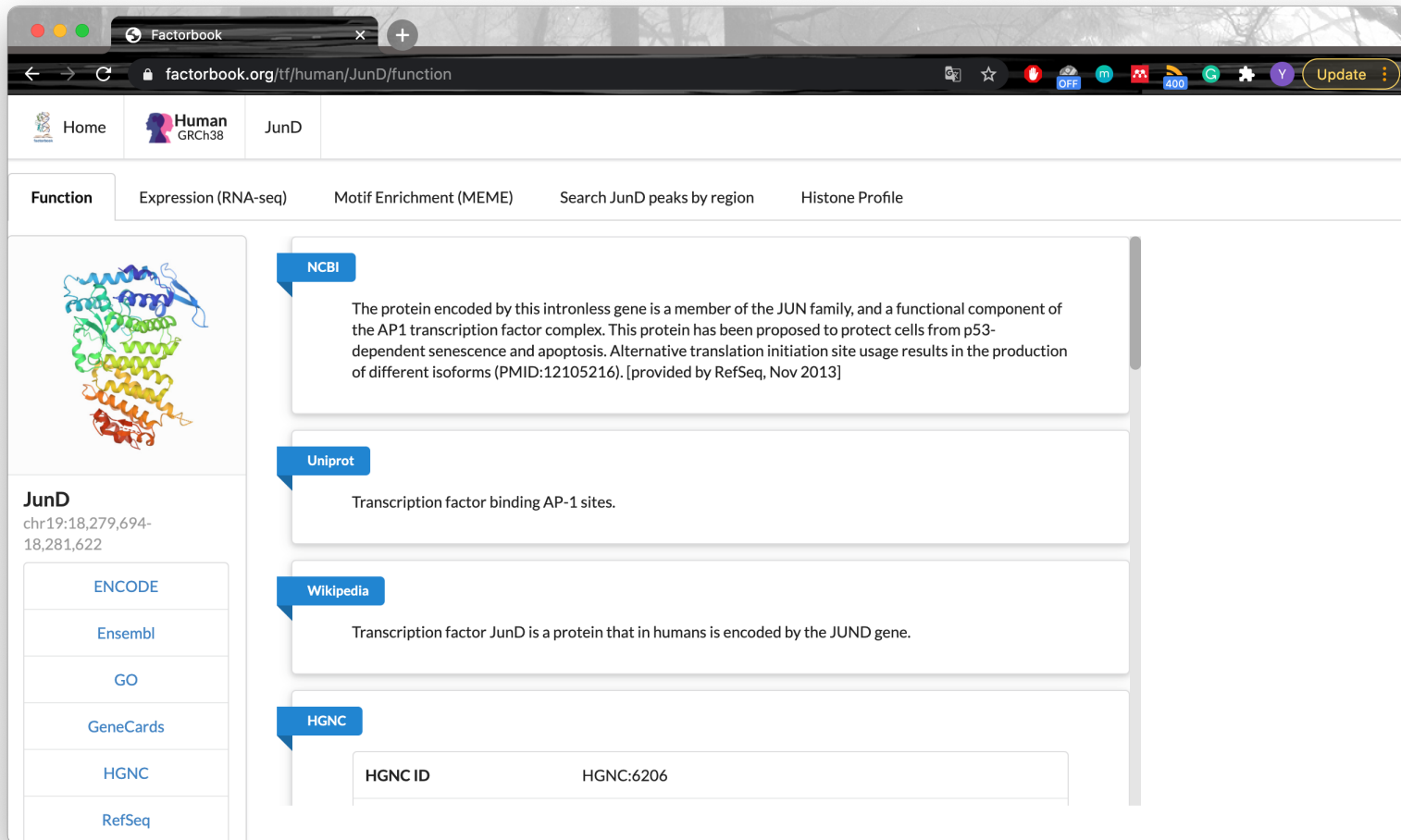
<http://jaspar.genereg.net>



Database Demo: FactorBook

- The example we use here is JunD.

<http://www.factorbook.org/mediawiki/index.php/JunD>



The screenshot displays the FactorBook website interface for the JunD protein. The browser address bar shows the URL `factorbook.org/tf/human/JunD/function`. The page features a navigation bar with links to Home, Human GRCh38, and JunD. Below this, a tabbed interface is visible, with the 'Function' tab selected. On the left side, there is a 3D ribbon diagram of the JunD protein structure. Below the diagram, the protein name 'JunD' is listed along with its genomic coordinates: chr19:18,279,694-18,281,622. A vertical list of database links is provided, including ENCODE, Ensembl, GO, GeneCards, HGNC, and RefSeq. The main content area on the right contains several information boxes. The 'NCBI' box provides a detailed description of the protein as a member of the JUN family and a functional component of the AP1 transcription factor complex. The 'Uniprot' box states that the protein is a transcription factor binding AP-1 sites. The 'Wikipedia' box provides a general definition of JunD as a protein encoded by the JUND gene. The 'HGNC' box displays the HGNC ID as HGNC:6206.

Factorbook

factorbook.org/tf/human/JunD/function

Home Human GRCh38 JunD

Function Expression (RNA-seq) Motif Enrichment (MEME) Search JunD peaks by region Histone Profile

JunD
chr19:18,279,694-18,281,622

ENCODE
Ensembl
GO
GeneCards
HGNC
RefSeq

NCBI

The protein encoded by this intronless gene is a member of the JUN family, and a functional component of the AP1 transcription factor complex. This protein has been proposed to protect cells from p53-dependent senescence and apoptosis. Alternative translation initiation site usage results in the production of different isoforms (PMID:12105216). [provided by RefSeq, Nov 2013]

Uniprot

Transcription factor binding AP-1 sites.

Wikipedia

Transcription factor JunD is a protein that in humans is encoded by the JUND gene.

HGNC

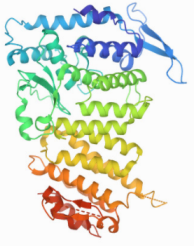
HGNC ID HGNC:6206

Database Demo: FactorBook

factorbook.org/tf/human/JunD/function

Home Human GRCh38 JunD

Function Expression (RNA-seq) **Motif Enrichment (MEME)** Search JunD peaks by region Histone Profile



JunD
chr19:18,279,694-18,281,622

[ENCODE](#)
[Ensembl](#)
[GO](#)
[GeneCards](#)
[HGNC](#)
[RefSeq](#)

14 experiments performed

[download](#)

Experiment Accession	Cell Type	Date Released ▲	Lab	Replicated Peak File Accession
ENCSR000BGK	HepG2	Jul 2011	Richard Myers	ENCFF539GRW
ENCSR000BKP	H1	Jul 2011	Richard Myers	ENCFF646IUA
ENCSR000EBZ	H1	Oct 2011	Michael Snyder	ENCFF443HNU
ENCSR000EEI	HepG2	Oct 2011	Michael Snyder	ENCFF430PEI
ENCSR000EGN	K562	Oct 2011	Michael Snyder	ENCFF213EYD

< 1 2 3 >

Database Demo: FactorBook

factorbook.org/tf/human/JunD/motif/ENCSR000DYS

Home Human GRCh38 JunD

Function Expression (RNA-seq) **Motif Enrichment (MEME)** Search JunD peaks by region Histone Profile

A549 1 exp

GM12878 1 exp

Michael Snyder (ENCSR000DYS)

H1 2 exps

HCT116 1 exp

HepG2 2 exps

K562 1 exp

MCF-7 1 exp

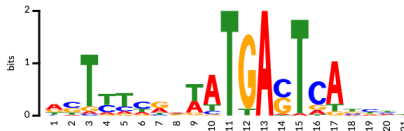
SK-N-SH 2 exps

T47D 1 exp

liver 2 exps

De novo motif discovery in GM12878 (ENCSR000DYS) by MEME

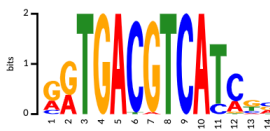
Download motifs from this dataset in MEME format



E-value	2e-276
Occurrences	2,803 / 7,452 peaks

Best external database match:
BATF_HUMAN.H11MO.0.A (HOCOMOCO)

export logo reverse complement show consensus sequence show more



E-value	1.5e-79
Occurrences	935 / 7,452 peaks

Best external database match:
MA0656.1/JDP2(var.2) (JASPAR)

Database Demo: JASPAR

- Go to JASPAR database, and search JunD again:

The screenshot shows the JASPAR 2020 database interface. The search bar contains 'JunD' and the search button is labeled 'Search Q'. Below the search bar, it says 'Examples: SPI1, P17676, ChIP-seq, Homo sapiens' and 'Advanced Options'. The results section shows '10 profile(s) found' and a table of search results. The first row is highlighted with a red box. The table has columns: ID, Name, Species, Class, Family, and Logo. The first row is MA0491.1, JUND, Homo sapiens, Basic leucine zipper factors (bZIP), Jun-related factors, and a logo showing the sequence TGAATCAAT. The right sidebar shows 'Analyze selected profiles' and a message: 'Please select matrix profiles on the left side to add to your cart or perform the following analysis.' There is an 'Add to cart' button and a notification that says 'You have 0 profile(s) in your cart. You can add profiles to the cart to download or perform analysis.' There are also 'Add to cart' and 'View cart' buttons at the bottom of the sidebar.

JASPAR²⁰²⁰

Search profile(s)

JunD

Examples: SPI1, P17676, ChIP-seq, Homo sapiens

Search Q

Advanced Options

10 profile(s) found

Display 10 profiles

Filter:

ID	Name	Species	Class	Family	Logo
MA0491.1	JUND	Homo sapiens	Basic leucine zipper factors (bZIP)	Jun-related factors	
MA0491.2	JUND	Homo sapiens	Basic leucine zipper factors (bZIP)	Jun-related factors	
MA0492.1	JUND(var.2)	Homo sapiens	Basic leucine zipper factors (bZIP)	Jun-related factors	
MA1141.1	FOS::JUND	Homo sapiens	Basic leucine zipper factors (bZIP)::Basic	Fos-related factors::Jun-related	

Analyze selected profiles –

Please select matrix profiles on the left side to add to your cart or perform the following analysis.

Add to cart

You have 0 profile(s) in your cart. You can add profiles to the cart to download or perform analysis.

Add to cart View cart

Database Demo: JASPAR

- Summary page for JUND transcription factor

JASPAR2020

Cart 0JASPAR Blog

Home

About

Search

Browse JASPAR CORE

Unvalidated Profiles

Browse Collections

Tools

RESTful API

Download Data

Matrix Clusters

Genome Tracks

Detailed information of matrix profile **MA0491.1**

Home > Matrix > MA0491.1

Profile summary

Add

Name: JUND

Matrix ID: MA0491.1

Class: Basic leucine zipper factors (bZIP)

Family: Jun-related factors

Collection: CORE

Taxon: Vertebrates

Species: [Homo sapiens](#)

Data Type: ChIP-seq

Validation: [21526160](#)

Uniprot ID: [P17535](#)

Source: [ENCODE](#)

Comment:

Sequence logo

Download SVG



Frequency matrix

JASPARTRANSFACMEMERAW PFM

Reverse comp.

A [	12389	15163	0	0	38710	0	0	1547	38710	0	5786	]
C [	759	4824	0	0	0	22857	0	37163	0	5136	14956	]
G [	13640	18723	0	38710	0	14078	0	0	0	2301	12429	]
T [	11922	0	38710	0	0	1775	38710	0	0	31273	5539	]

Demonstration on MEME

- Motif-based sequence analysis tools

← → ↻ 🔒 meme-suite.org/meme/index.html

The MEME Suite

Motif-based sequence analysis tools

MEME Suite 5.3.3

- Motif Discovery
- Motif Enrichment
- Motif Scanning
- Motif Comparison
- Gene Regulation
- Manual
- Guides & Tutorials
- Sample Outputs
- File Format Reference
- Databases
- Download & Install
- Help
- Alternate Servers
- Authors & Citing
- Recent Jobs
- ◀ Previous version 5.3.2

Workflow:

- Input:** Your DNA, RNA or protein sequences; Motif databases; Your DNA, RNA or protein motifs; GO databases
- Motif Discovery:** MEME, STREME, MEME-ChIP, GLAM2, MoMo → Discovered motifs (de novo)
- Motif Enrichment:** CentriMo, AME, SpaMo, GOMo → Enriched motifs
- Annotated motifs:** GO function, GO compartment, GO process
- Motif Scanning:** FIMO, MAST, MCAST, GLAM2SCAN → Annotated sequences
- Motif Comparison:** Tomtom → Aligned motifs
- T-Gene:** Your BED file of genomic loci → Regulatory gene targets

Tools:

- MEME:** Multiple Em for Motif Elicitation
- CentriMo:** Local Motif Enrichment Analysis
- FIMO:** Find Individual Motif Occurrences
- STREME:** Sensitive, Thorough, Rapid, Enriched Motif Elicitation
- AME:** Analysis of Motif Enrichment
- MAST:** Motif Alignment & Search Tool
- MEME-ChIP:** Motif Analysis of Large Nucleotide Datasets
- SpaMo:** Spaced Motif Analysis Tool
- MCAST:** Motif Cluster Alignment and Search Tool
- GLAM2:** Gapped Local Alignment of Motifs
- GOMo:** Gene Ontology for Motifs
- GLAM2Scan:** Scanning with Gapped Motifs
- MoMo:** Modification Motifs
- Tomtom:** Motif Comparison Tool
- GT-Scan:** Identifying Unique Genomic Targets
- T-Gene:** Predicting Target Genes
- DREME:** Discriminative Regular Expression Motif Elicitation

Demonstration on MEME

- Web tools for different proposes:
 - Discover motifs: MEME, DREAM, GLAM2
 - Search databases: MAST, FIMO, MCAST, GLAM2SCAN
 - Compare motifs: TOMTOM



Web tool demo: MEME



Version 4.9.1

Use this form to submit DNA or protein sequences to MEME. MEME will analyze your sequences for similarities among them and produce a description (**motif**) for each pattern it discovers.

- Input sequences in FASTA format

the **actual sequences** here ([Sample Protein Input Sequences](#)):

```
>At1g01140.1_4-2-4_SnRK3.12 SNF1-related Protein Kinase,
subfamily 3
MSGSRRKATPASRTRVGNYEMGRTLGECSFAKVKYAKNTVTGDQ
AAIKILDREKVFRHKM
VEQLKREISTMKLIKHPNVVEIIEVMASKTKIYVLELVNGGELFDKIAQ
QGRLKEDEAR
```

- Submit query

☐ Search given **strand** only
☐ Look for **palindromes** only

Version 4.9.1

Please send comments and questions to: meme@nbcn.net

Powered by [Opal](#)

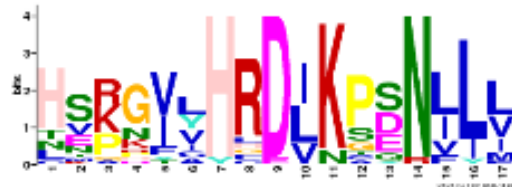
Web tool demo: MEME

DISCOVERED MOTIFS

Motif Overview

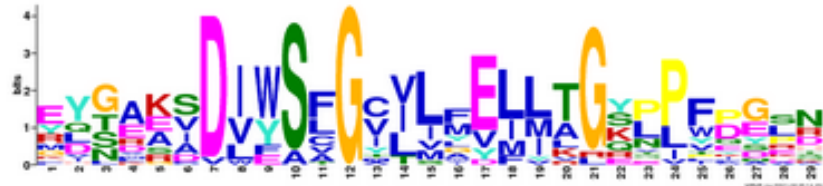
[Motif 1](#)

- $1.8e-100$
- 19 sites



[Motif 2](#)

- $6.2e-121$
- 19 sites

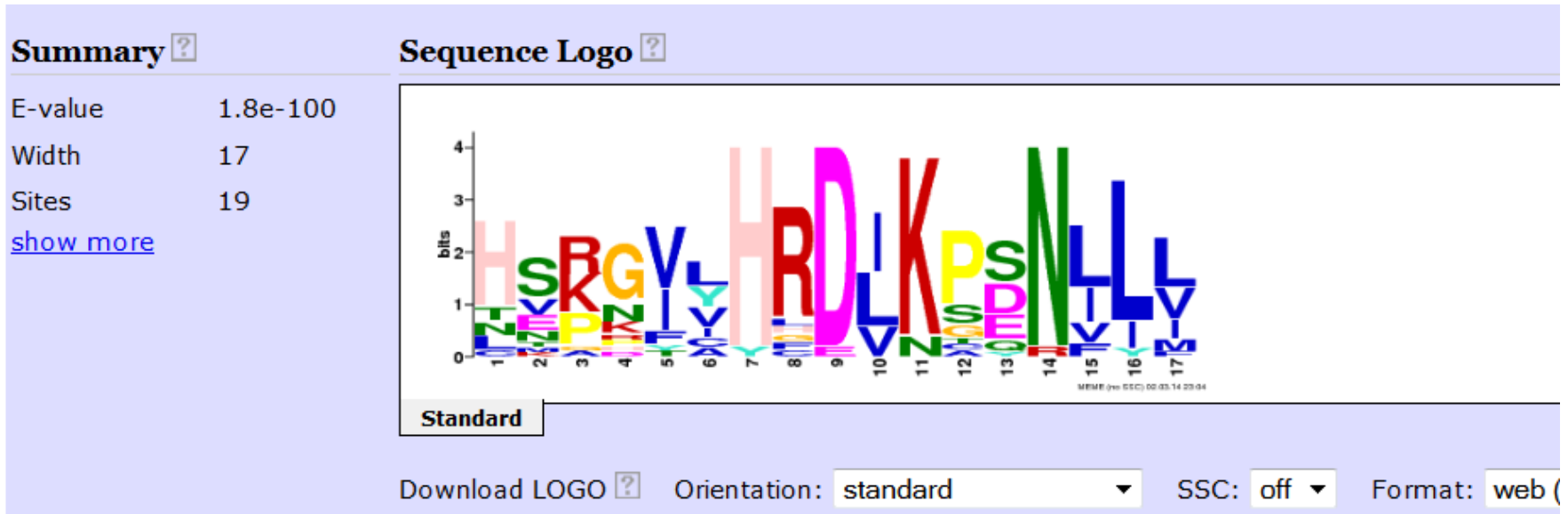


[Motif 3](#)

- $1.5e-071$
- 9 sites



Web tool demo: MEME



Regular expression [?](#)

HS[KRP]G[VI][LVY]HRD[ILV]KP[SDE]N[LI]L[LV]

Further Analysis [?](#)

Submit this motif to [?](#) [?](#) [?](#) Mouse-over buttons for more information.

Data Formats [?](#)

View the motif in ☐ PSPM Format [?](#) ☐ PSSM Format [?](#) ☐ BLOCKS Format [?](#) ☐ FASTA Format [?](#)

Check list

- What are the three important features of a motif (or domain)?
- What are the five different representations of a motif?
- Name a database and corresponding data related to motif.
- What is motif discovery problem?

Extra Exercise

- Suppose in the genomic background, the percentages of A, C, G and T are 30%, 20%, 20% and 30%, respectively. What is the probability for each the following patterns to appear on either strand of a random DNA sequence of 5bp long, assuming each position is independent of the others? Explain how these values are computed.
- AACTT
- ACNG, where N can be any of the four nucleotides
- ATA