

# Scoring matrix –BLOSUM 62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	+2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

# Overview of homology search strategy

## 2.) Which sequence should I use as the query?

- Protein
- cDNA
- Genomic

# Overview of homology search strategy

## 2.) Which sequence should I use as the query?

### cDNA (BlastN)

Sequences producing significant alignments:			Score (bits)	E Value
<a href="#">gnl dmel FBtr0082091</a>	type=mRNA; loc=3R:complement(5531512.....		38	0.87
<a href="#">gnl dmel FBtr0085316</a>	type=mRNA; loc=3R:complement(24562831....		38	0.87
<a href="#">gnl dmel FBtr0071092</a>	type=mRNA; loc=X:7757325..7762681; nam...		36	3.4
<a href="#">gnl dmel FBtr0085763</a>	type=mRNA; loc=3R:27088887..27089539; ...		36	3.4
<a href="#">gnl dmel FBtr0087330</a>	type=mRNA; loc=2R:11021527..11023229; ...		36	3.4
<a href="#">gnl dmel FBtr0079508</a>	type=mRNA; loc=2L:complement(7717052.....		36	3.4
<a href="#">gnl dmel FBtr0079312</a>	type=mRNA; loc=2L:complement(6686819.....		36	3.4

### Protein (TblastN)

Sequences producing significant alignments:			Score (bits)	E Value
<a href="#">gnl dmel FBtr0086108</a>	type=mRNA; loc=2R:2160554..2164644; na...		53	3e-07
<a href="#">gnl dmel FBtr0088077</a>	type=mRNA; loc=2R:7195380..7204666; na...		47	1e-05
<a href="#">gnl dmel FBtr0076455</a>	type=mRNA; loc=3L:9378742..9380127; na...		28	9.2

# Overview of homology search strategy

## 2.) Which sequence should I use as the query?

### Protein v.s cDNA

query: S	A	L	target: S	A	L
query: TCT	GCA	TTG	target: AGC	GCT	CTA
			<u>Base level identity</u>		
			Protein: ~ 5%		
			Nucleotide: ~ 25%		

Protein: 100%  
Nucleotide: 33%

**Searching at the protein level is much more sensitive**

# Overview of homology search strategy

## 2.) Which sequence should I use as the query?

If you want to identify similar feature at the DNA level. Be Cautious with genomic sequence initiated search

- **Low complexity region**
- **repeats**

# Overview of homology search strategy

## 5.) How to optimize the search ?

- Scoring matrices
- Gap penalty
- Expectation / cut off

Example

# Overview of homology search strategy

## 8.) How to align (compare) my query and the hits ?

- Global alignment
- Local alignment

ClustalW/ClustalX

# Limitations of Generic Scoring Matrix

- ❖ Short specific pattern
- ❖ Position specific information.

# Binary patterns for protein and DNA

## Examples:

- Caspase recognition site:

**[EDQN] X [^RKH] D [ASP]**

Observe: Search for potential caspase recognition sites with BaGua

# Why BLAST may fail to find the real orthologue?

**Position –specific information about conserved domains is  
IGNORED in single sequence –initiated search**

BID_MOUSE	SESQEEIHN	IARHLAQIGDEM	DHNIQPTLVR
BAD_MOUSE	APPNLWAAQR	YGRELRRMSDEF	EGSFKGLPRP
BAK_MOUSE	PLEPNSILGQ	VGRQLALIGDDI	NRRYDTEFQN
BAXB_HUMAN	PVPQDASTKK	LSECLKRIGDEL	DSNMELQRFMI
BimS	EPEDLRPEIR	IAQELRRIGDEF	NETYTRRVFA
HRK_HUMAN	LGLRSSAAQL	TAARLKALGDEL	HQRTMWRRRA
Eg1-1	DSEISSIGYE	IGSKLAAMCDDF	DAQMMSYSAH

BID_MOUSE	SESQEEIHN	IARHLAQIGDEM	DHNIQPTLVR
sequence X	SESSSELLHN	SAGHAAQLFDSM	RLDIGSTAHR
sequence Y	PGLKSSAANI	LSQQLKGIGDDL	HQRMMSYSAH

## Basic concept of motif identification 2.

How do we represent the position specific preference ?

BID_MOUSE	I	A	R	H	L	A	Q	I	G	D	E	M
BAD_MOUSE	Y	G	R	E	L	R	R	M	S	D	E	F
BAK_MOUSE	V	G	R	Q	L	A	L	I	G	D	D	I
BAXB_HUMAN	L	S	E	C	L	K	R	I	G	D	E	L
BimS	I	A	Q	E	L	R	R	I	G	D	E	F
HRK_HUMAN	T	A	A	R	L	K	A	L	G	D	E	L
Egl-1	I	G	S	K	L	A	A	M	C	D	D	F

Binary pattern:

[HEQCRK]

L

[GSC]

X

[^ILMFV]

## Basic concept of motif identification 2.

### How do we represent the position specific preference ?

BID_MOUSE	I	A	R	H	L	A	Q	I	G	D	E	M
BAD_MOUSE	Y	G	R	E	L	R	R	M	S	D	E	F
BAK_MOUSE	V	G	R	Q	L	A	L	I	G	D	D	I
BAXB_HUMAN	L	S	E	C	L	K	R	I	G	D	E	L
BimS	I	A	Q	E	L	R	R	I	G	D	E	F
HRK_HUMAN	T	A	A	R	L	K	A	L	G	D	E	L
Egl-1	I	G	S	K	L	A	A	M	C	D	D	F

Statistical  
representation

G: 5 -> 71%

S: 1 -> 14 %

C: 1 -> 14 %

# Protein motif /domain

- Structural unit
- Functional unit
- Signature of protein family

**How are they defined?**

# Practice: Using MEME to identify motifs shared by a set of proteins

1. Load the sequence file to MEME.
2. Make sure you input your email address for results.

# Representation of positional information in specific motif

Binary patterns:

M-C-N-S-S-C-[MV]-G-G-M-N-R-R.

Positional matrix:

-2.499	-2.269	-5.001	-4.568	-2.418	-4.589	-3.879	1.971	-4.330	1.477	-1.241	-4.221	-4.590	-4.097	-4.293	-3.808	0.495	2.545	-3.990
1.627	-2.453	-1.804	-1.746	-3.528	2.539	1.544	-3.362	-1.440	-3.391	-2.490	-1.435	-3.076	-1.571	0.501	0.201	-1.930	-2.707	-3.990
-1.346	-2.872	-1.367	0.699	-2.938	-2.427	-0.936	-2.632	-0.095	1.147	-1.684	-1.111	-2.531	1.174	2.105	1.057	-1.400	-2.255	-2.990
1.045	1.754	-1.169	-0.576	-2.756	-2.212	1.686	-2.576	0.951	-2.438	-1.544	0.857	-2.301	1.891	1.556	-1.097	-1.180	-2.155	-2.990
-3.385	-2.965	-5.039	-4.313	-1.529	-5.006	-3.577	0.429	-4.094	3.154	-0.121	-4.440	-4.199	-3.292	-3.662	-4.198	-3.281	-1.505	-3.990
2.368	-3.197	-2.285	-1.533	-3.721	-2.945	1.815	-3.235	0.067	-3.061	-2.259	-1.680	-3.231	1.195	2.287	-2.009	-2.044	-2.825	-3.990
1.046	1.742	0.576	-0.734	-2.072	-2.234	-0.851	0.436	-0.548	-0.129	-0.974	-1.039	-2.318	2.368	0.667	-1.135	-1.076	-1.304	-2.990
0.715	-1.778	-3.820	-3.359	-1.535	-3.463	-2.571	3.060	-3.008	0.262	1.566	-2.996	-3.575	0.450	-3.001	-2.571	-1.765	0.193	-2.990
-2.053	0.965	-2.767	-3.509	-4.520	3.654	-3.242	-4.548	-3.338	-4.968	-3.789	-2.462	-4.048	-3.738	-3.266	-2.596	-3.573	-3.990	-3.990

# Scoring sequence based on Model

Seq: A S L D E L G D E



1 2 3 4

A 1 0-4 .

C 2 5-1 .

D 1-3 9 .

... . . . .

score @ position\_1 =  
 $s(A/1) + s(S/2) + s(L/3) +$   
 $s(D / 4)$

An example of position specific matrix

Observe: Search for proteins  
encodes IL6 motif#2 in Dm  
genome.

# Practice: motif analysis of protein sequence using ScanProsite and Pfam

1. Open two tabs for Pfam, input one of the Blast hits and one candidate TNF to each data window.
2. Compare the results