Standalone Blast - Basis

```
Database: dbs/dm6.44.fasta
           17,874 sequences; 102,739,733 total letters
```

```
Query= sp|P08505|IL6_MOUSE Interleukin-6 OS=Mus musculus OX=10090 GN=Il6
PE=1 SV=1

Length=211

Sequences producing significant alignments:                          Score     E
                                                                     (Bits)  Value

  FBgn0046706 type=gene; loc=2R:4014111..4049342; ID=FBgn0046706;...  30.0     4.0
```

```
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Neighboring words threshold: 13
Window for multiple hits: 40
```

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

1.) Which sequence should I use as the query?

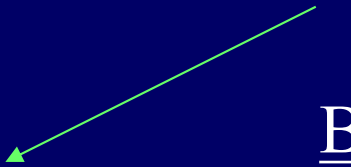
- Protein
- cDNA
- Genomic

Overview of homology search strategy

1.) 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
```



	Base level identity
Protein: 100%	Protein: ~ 5%
Nucleotide: 33%	Nucleotide: ~ 25%

Searching at the protein level is much more sensitive

Overview of homology search strategy

1.) Which sequence should I use as the query?

cDNA (BlastN)

```
                                                                    Score     E
Sequences producing significant alignments:                         (bits)  Value

gnl|dmel|FBtr0082091  type=mRNA; loc=3R:complement(5531512.....      38    0.87
gnl|dmel|FBtr0085316  type=mRNA; loc=3R:complement(24562831....      38    0.87
gnl|dmel|FBtr0071092  type=mRNA; loc=X:7757325..7762681; nam...      36    3.4
gnl|dmel|FBtr0085763  type=mRNA; loc=3R:27088887..27089539; ...      36    3.4
gnl|dmel|FBtr0087330  type=mRNA; loc=2R:11021527..11023229; ...      36    3.4
gnl|dmel|FBtr0079508  type=mRNA; loc=2L:complement(7717052.....      36    3.4
gnl|dmel|FBtr0079312  type=mRNA; loc=2L:complement(6686819.....      36    3.4
```

Protein (TblastN)

```
                                                                    Score     E
Sequences producing significant alignments:                         (bits)  Value

gnl|dmel|FBtr0086108  type=mRNA; loc=2R:2160554..2164644; na...      53    3e-07
gnl|dmel|FBtr0088077  type=mRNA; loc=2R:7195380..7204666; na...      47    1e-05
gnl|dmel|FBtr0076455  type=mRNA; loc=3L:9378742..9380127; na...      28    9.2
```

Overview of homology search strategy

2.) How should I customize the BLAST search to suit the objectives of my project?

- Which substitution matrix to use?
- How to adjust gap penalties?

BLOSUM - Blocks Substitution Matrices

-- Clustering threshold

BLOSUM 90 – Blocks with >= **90%** identity are counted as one to compute the substitution score

⋮

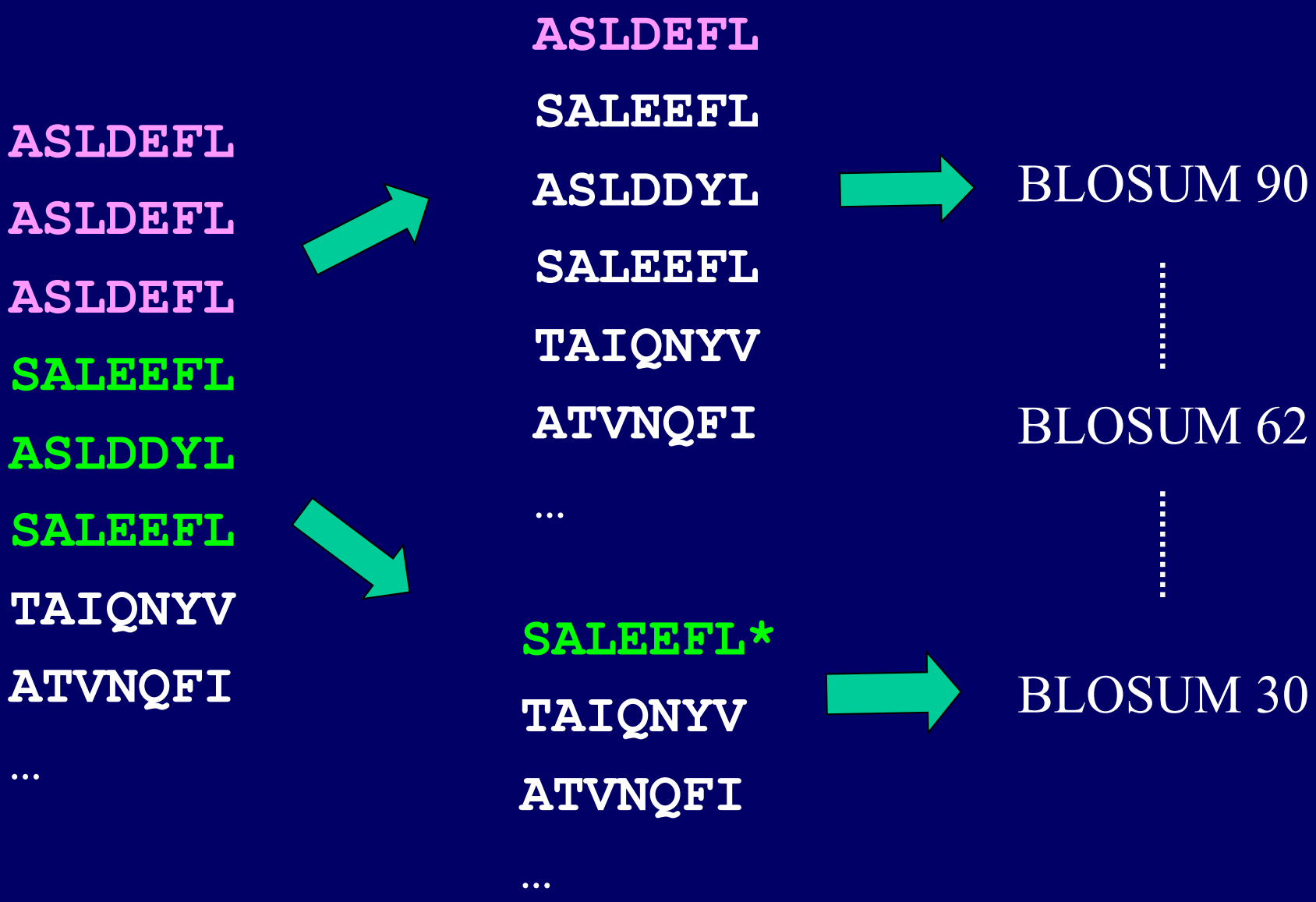
BLOSUM 62

⋮

BLOSUM 30 – Blocks with >= **30 %** identity are counted as one to compute the substitution score

BLOSUM - Blocks Substitution Matrices

-- Clustering threshold



Comparison of Blosum matrixes

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F
L	-2	-3	-4	-5	-2	-3	-4	-5	-4	1	5	-3	2	0

Blosum 90

L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0

Blosum 62

L	-1	-2	-2	-1	0	-2	-1	-2	-1	2	4	-2	2	2

Blosum 30

Why BLAST uses BLOSUM62 as the Default.

Proc Natl Acad Sci U S A. 1992 Nov 15; 89(22): 10915–10919.
doi: 10.1073/pnas.89.22.10915

Amino acid substitution matrices from protein blocks.

S Henikoff and J G Henikoff

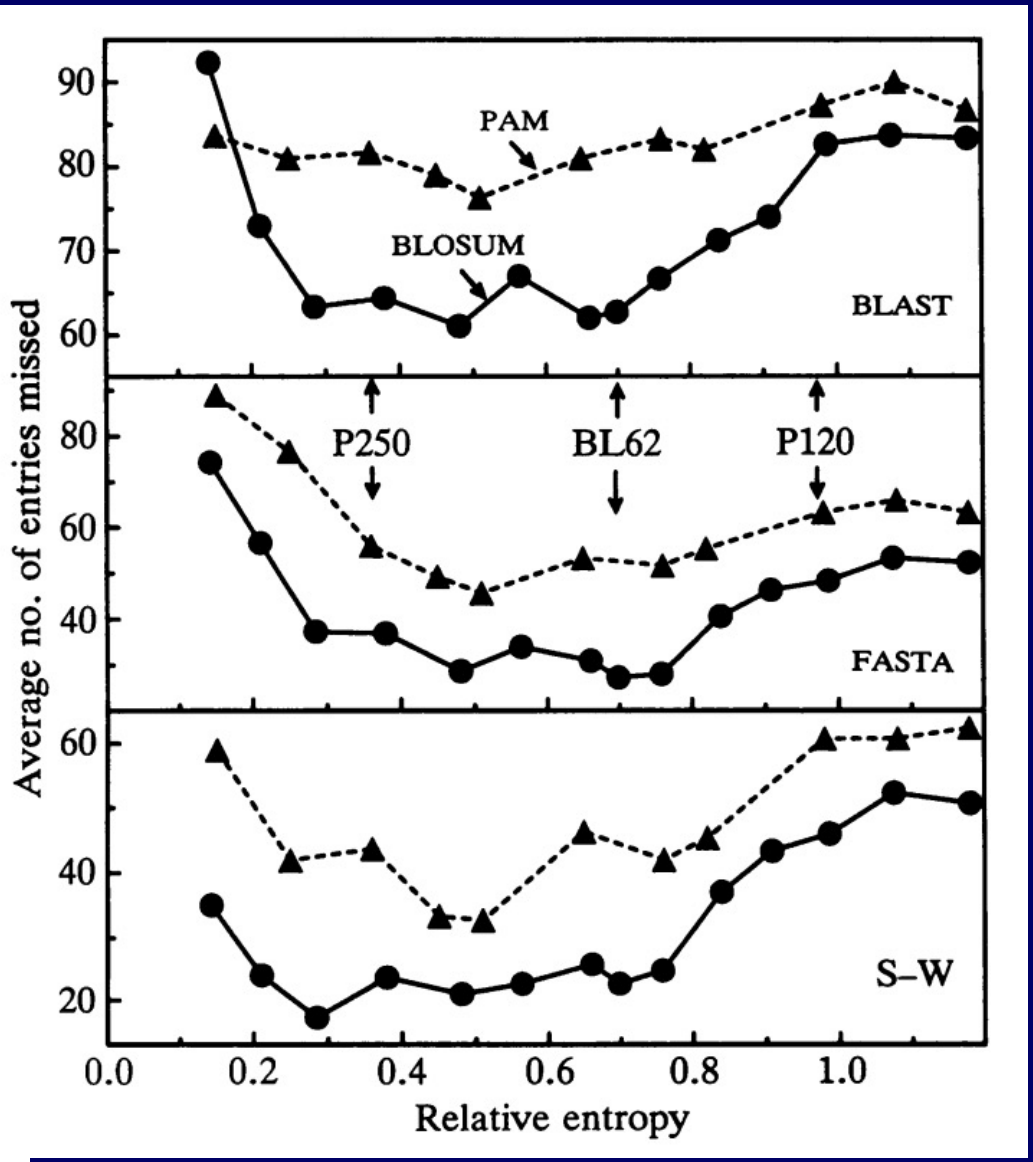


FIG. 3. Searching performance of programs using members of the guanine nucleotide-binding protein-coupled receptor family as queries and matrices from the BLOSUM and PAM series scaled in half-bits (11). Removal of this family from the BLOCKS data base led to a nearly identical matrix with similar performance. Matrices represented (left to right) are BLOSUM (BL) 30, 35, 40, 45, 50, 55, 60, 62, 65, 70, 75, 80, 85, and 90 and PAM (P) 400, 310, 250, 220, 200, 160, 150, 140, 120, 110, and 100. The average numbers of true positive Swiss-Prot entries missed are shown for LSHRRAT, RTARAT, and UL33$HCMVA versus Swiss-Prot 20. Results using BLAST and FASTA or SSEARCH (S–W) are not comparable to each other, since different detection criteria were used for the three programs.

Q: Which substitution matrix will you use to increase the sensitivity for identifying a distant ortholog?

a.) Blosum 40

b.) Blosum 60

c.) Blosum 90

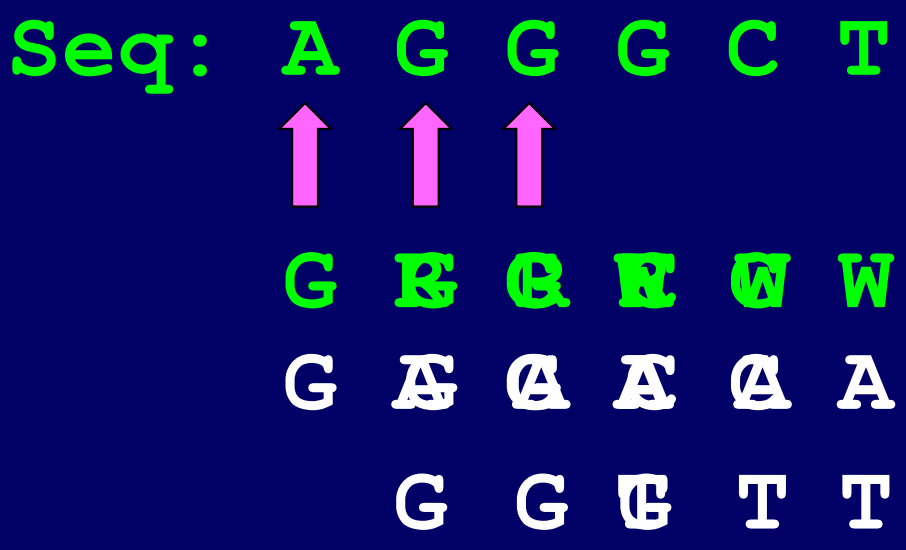
Limitations of Generic Scoring Matrix

- Short specific pattern
- Position specific information.

Specific patterns

1. DNA pattern – Transcription factor binding site.
2. Short protein pattern – enzyme recognition sites.
3. Protein motif/signature.

Searching for binary (string) patterns



Positive match

Binary patterns for protein and DNA

Examples:

- **Caspase recognition site:**

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

Observe: Search for potential caspase recognition sites with BaGua

Does binary pattern convey all the information ?

Why a BLAST match is refused by the family ?

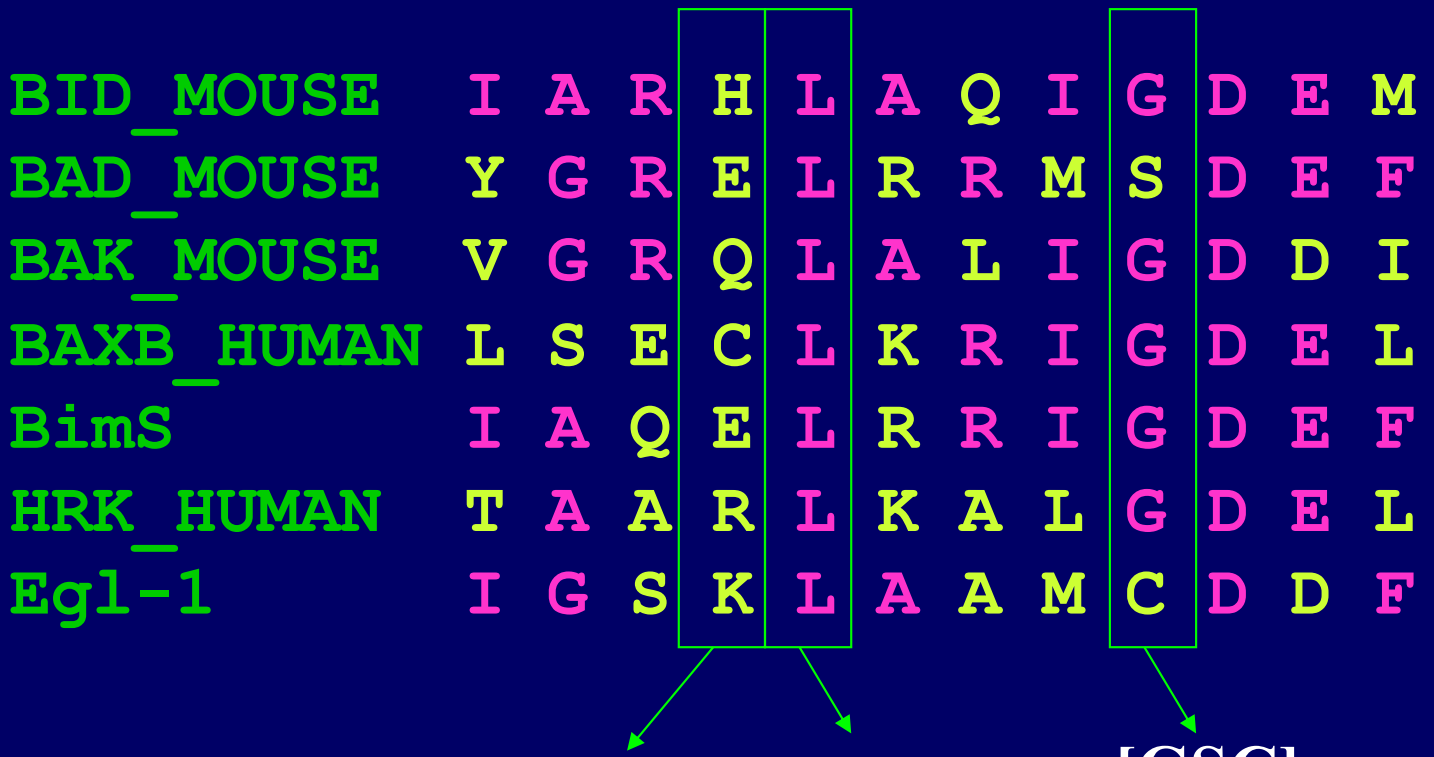
Position –specific information about conserved domains is IGNORED in single sequence –initiated search

```
BID_MOUSE     SESQEEIIHN IARHLAQIGDEM DHNIQPTLVR
BAD_MOUSE     APPNLWAAQR YGRELRRMSDEF EGSFKGLPRP
BAK_MOUSE     PLEPNSILGQ VGRQLALIGDDI NRRYDTEFQN
BAXB_HUMAN    PVPQDASTKK LSECLKRIGDEL DSNMELQRMI
BimS          EPEDLRPEIR IAQELRRIGDEF NETYTRRVFA
HRK_HUMAN     LGLRSSAAQL TAARLKALGDEL HQRTMWRRRA
Egl-1         DSEISSIGYE IGSKLAAMCDDF DAQMMSYSAH

BID_MOUSE     SESQEEIIHN 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 ?



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 %