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	

BLAST – Basic Local Alignment Search Tool

It is based on local alignment, -- highest score is the only priority in terms of finding alignment match.

-- determined by scoring matrix, gap penalty

It is **optimized** for searching large data set instead of finding the best alignment for two sequences

Global vs. Local Alignment

Before calculating the similarity score, we need an alignment --

Global alignment: from start to end

Local alignment: best matches on any segment of participating sequence.

Practice : try local alignment and global alignment of the same pair of sequence

- Start two new browser tabs with the alignment server.
- Open the test sequence file, copy seq1 and seq2 to the respective windows in the alignment web page.
- Select Blosum62 as the scoring matrix on both pages.
- Run one set for Local alignment, the other for global alignment.

Global Alignment seq1 & seq2

```
Score: -57 at (seq1)[1..90] : (seq2)[1..92]

>seq1   MA-----STVTSCLEPTEVFMDLWPEDHSNWQELSPLEPSDPLNPPTPPRAAPSPVVPST
        :.     ::. .     :  . :  . .  .. :  .   .  :    : :. : :  ..
>seq2   MSHGIQMSTIKKRRSTDEEVFCLPIKGREIYEILVKIYQIENYNMECAPPAGASSVSVGA

>seq1   EDYGGDFDFRVGFVEAGTAKSVTCTYSPVLNKVYC
         .      :   . :    .:     ::.  . .
>seq2   TEAEPTEVFMDLWPED---HSNWQELSPLEPSDHM
```

14 identical matches

Local Alignment seq1 & seq2 with BLOSUM 62

```
Score: 156 at (seq1)[10..36] : (seq2)[64..90]

    10 EPTEVFMDLWPEDHSNWQELSPLEPSD
       ||||||||||||||||||||||||||||
    64 EPTEVFMDLWPEDHSNWQELSPLEPSD
```

27 identical matches

Finding the best alignment = Get the highest score

The consideration on whether to open/extend a gap is weighed by its effect on the **total score** of the alignment.

Optimization - Dynamic programming

Global vs. Local Alignment

Q:

Can a Global alignment produce a Local alignment ?

Can a Local alignment produce a Global alignment ?

BLAST – Basic Local Alignment Search Tool

1. **A high similarity core (2-4aa)**
2. **Often without gap**

```
Query: M A T W L I .
Word : M A T
         A T W
           T W L
             W L I
```

1. **For each word, find matches with Score > T.**
2. **Extend the match as long as profitable.**

   - High Scoring segment Pair (best local alignment)
3. **Find the P and E value for HSP(s) with Score > cut off*.**

* Cut off value can be automatically calculated based on E

BLAST – Basic Local Alignment Search Tool

The P and E value for HSP(s) : based on the **total score (S)** of the identified "best" local alignment.

P (S) : the probability that two random sequences, one the length of the query and the other the entire length of the database, could achieve the score S.

E (S) : The expectation of observing a score >= S in the target database.

For a given database, there is a one to one correspondence between S and E(s) -- choosing E determines cut off score

BLAST – Basic Local Alignment Search Tool

BLASTN
BLASTP
TBLASTN

compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.

BLASTX

compares a nucleotide query sequence translated in all reading frames against a protein sequence database

TBLASTX

compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Please note that tblastx program cannot be used with the nr database on the BLAST Web page.

BLAST – Advanced options : all adjustable in stand alone BLAST

-F Filter query sequence [String] default = T

-M Matrix [String] default = BLOSUM62

-G Cost to open gap [Integer] default = 5 for nucleotides 11 proteins

-E Cost to extend gap [Integer] default = 2 nucleotides 1 proteins

-q Penalty for nucleotide mismatch [Integer] default = -3

-r reward for nucleotide match [Integer] default = 1

-e expect value [Real] default = 10

-W wordsize [Integer] default = 11 nucleotides 3 proteins

-T Produce HTML output [T/F] default = F

Overview of homology search strategy

1.) Where should I search?

- **NCBI**

  Has pretty much every thing that has been available for some time

- **Genome projects**

  Has the updated information (DNA sequence as well as analysis result)

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
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)

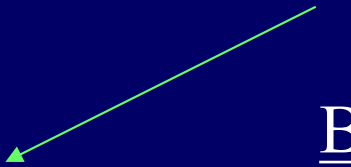
Sequences producing significant alignments:		Score (bits)	E Value
gnl|dmel|FBtr0086108	type=mRNA; loc=2R:2160554..2164644; na...	53	3e-07
gnl|dmel|FBtr0088077	type=mRNA; loc=2R:7195380..7204666; na...	47	1e-05
gnl|dmel|FBtr0076455	type=mRNA; loc=3L:9378742..9380127; na...	28	9.2

Overview of homology search strategy

2.) 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

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

3.) Which program to use?

1. Smith-Waterman vs. Blast.
2. Different flavors of BLAST

Overview of homology search strategy

4.) Which data set should I search?

- **Protein sequence (known and predicted)**
  blast**P, Smith_Waterman**
- **Genomic sequence**
  Tblast**N**
- **EST**
  Tblast**N**
- **Predicted genes**
  Tblast**N**

Overview of homology search strategy

5.) How to optimize the search ?

- Scoring matrices
- Gap penalty
- Expectation / cut off

Example

Overview of homology search strategy

6.) How do I judge the significance of the match ?

- P-value, E -value
- Alignment
- Structural / Function information

Overview of homology search strategy

7.) How do I retrieve related information about the hit(s) ?

- **NCBI is relatively easy**

  The scope of information collection can be enlarged by searching (linking) multiple databases (links). example

- **Genome projects often have their own interface and logistics (ie. Ensemble, wormbase, MGI, etc.)**

Overview of homology search strategy

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

- **Global alignment**
- **Local alignment**

ClustalW/ClustalX