# Blast output

```
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 (Bits) | E Value |
|---|---|---|
| FBgn0046706 type=gene; loc=2R:4014111..4049342; ID=FBgn0046706;... | 30.0 | 4.0 |

# Questions after the Blast search

**Questions:**

- How are the hits identified?

---

- What is the meaning of the score?

# Blast output

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

# Questions after the Blast search

**Questions:**

- How are the hits identified?

---

- What is the meaning of the score?

Observe & Practice: Scoring the similarity between two sequences.

# How to measure the similarity between two sequences

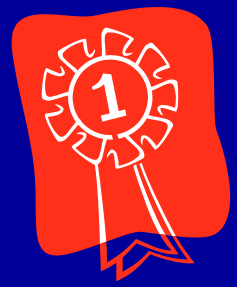
Q: which one is a better match to the query ?

```
Query:   M A T W L
Seq_A:   M A T P P
Seq_B:   M P P W I
```

# Judging the match using "Scoring Matrix"

Q: which one is a better match to the query ?

| Query: M A T W L | Query: M A T W L |
|---|---|
| Seq_A: M A T P P | Seq_B: M P P W I |
| Score: 5 4 5 -4 -3 | Score: 5-1-1 11 2 |
| Total: 7 | Total: 16 |



# “Scoring Matrix” assigns a score to each pair of amino acids

| | A | S | T | L | I | V | K | D | ... |
|---|---|---|---|---|---|---|---|---|---|
| L | -1 | -2 | -2 | 4 | 3 | 1 | -2 | -4 | |

**BLOSUM-62**

# BLOSUM - Blocks Substitution Matrices

**Block**: very well conserved region of a protein family. – perform the same (similar) function.

```
ASLDEFL
SALEDFL
ASLDDYL
ASIDEFY
ASIDEFY
...
```

```
AA: 6
AS: 3
SS: 0
```

$$\text{Score}(a1/a2) = 2 * \log 2 \frac{\text{observed frequency of } a1/a2}{\text{predicated frequency of } a1/a2}$$

# BLOSUM - Blocks Substitution Matrices

Block: very well conserved region of a protein family. – perform the same (similar) function.

**ASLDEFL**
**ASLEDFL**
**ASLDDYL**
**SALEEFL**
**ASLDDYL**
**SALEEFL**
...

Score (a1/a2)

- $> 0$: **observed frequency of a1/a2** $>$ **predicated frequency of a1/a2**
- $= 0$
- $< 0$: **observed frequency of a1/a2** $<$ **predicated frequency of a1/a2**

# BLOSUM - Blocks Substitution Matrices

Block: very well conserved region of a protein family. – perform the same (similar) function.

```
ASLDEFL
ASLEDFL
ASLDDYL
SALEEFL
ASLDDYL
SALEEFL
...
```

**observed frequency of L / I**
i.e: 0.03

>

**predicated frequency of L / I**
i.e: 0.1*0.1 = 0.01

Score (L/I) > 0

**Substitution of L / I is common in conserved sequences**

# BLOSUM - Blocks Substitution Matrices

Block: very well conserved region of a protein family. – perform the same (similar) function.

```
ASLDEFL
ASLEDFL
ASLDDYL
SALEEFL
ASLDDYL
SALEEFL
...
```

**observed frequency of L / K** i.e: 0.0002 **<** **predicated frequency of L / K** i.e: 0.1*0.1 = 0.01

Score (L/K) < 0

**Substitution of L / K is rare in conserved sequences**

# “Scoring Matrix” assigns a score to each pair of amino acids

| | A | S | T | L | I | V | K | D | ... |
|---|---|---|---|---|---|---|---|---|---|
| L | -1 | -2 | -2 | 4 | 3 | 1 | -2 | -4 | |

BLOSUM-62

# 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 | |

# 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

# Effect of matrices on Local Alignment

Observe: effect of matrices on the outcome of local alignment

**First name initial > L -- Align seq1 and seq 2 with "blosum62"**

**Others -- Align seq1 and seq 2 with "blosum35"**

# Effect of matrices on Local Alignment

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

**Blosum 62:**

**P / H: -2**

**L/M: 2**

```
Score: 206 at (seq1)[10..38] :
(seq2)[64..92]
10 EPTEVFMDLWPEDHSNWQELSPLEPSDPL
   |||||||||||||||||||||||||| .
64 EPTEVFMDLWPEDHSNWQELSPLEPSDHM
```

**Blosum 35:**

**P / H: -1**

**L/M: 3**

# Introducing a gap

```
Q: M A T W L I .
A: M A W T V A .

Scr: 5 4 -2 -2 1 -1

Total: 5
```

```
Q: M A T W I .
A: M A - W V .

Scr: 5 4 -? 11 3

Total = 22 - ?
```

Blosum 62:

Gap openning: -6 ~ -15

Gap Extension: -2 ~ -6

# Effect of gap penalty on Local Alignment

## Practice : effect of gap penalty on local alignment

**Set matrix to “blosum62”**

**Column 1,3,5, align seq1 and seq2 with “gap=15, ext=3,”**

**Column 2 and 4, align seq1 and seq2 with “gap=5, ext=1”**

# Effect of gap penalty on Local Alignment

**Blosum 62**

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

**Gap: -15**

**Ex: -3**

**Gap: -5 Ex: - 1**

```
Score: 161 at (seq1)[2..36] : (seq2)[53..90]
2  ASTV----TSCLEPTEVFMDLWPEDHSNWQELSPLEPSD
   || |    |   ||||||||||||||||||||||||||||
53 ASSVSVGATEA-EPTEVFMDLWPEDHSNWQELSPLEPSD
```

# 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

# 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