BLAST Quick Start

Algorithm Basics
• Introduction
• Words & extensions

Search Basics
• Programs
• Databases
• Submit a search
• Interpret the results
• BLAST options
• Format options
• Examples
Basic Local Alignment Search Tool

Compare protein and nucleic acid sequences:
• in NCBI databases
• in local databases (standalone BLAST)
• to a single protein or nucleotide sequence (BLAST 2 Sequences, or pairwise BLAST)

• local alignments; isolated regions of similarity
• fast and sensitive
• breaks the query sequence into “words”
• word hits to database sequences extended in both directions
Global vs Local Alignment

Seq1: WHEREISWALTHERNOW (16aa)
Seq2: HEWASHEREBUTNOWISHERE (21aa)

Global

Seq1: 1 W--HERE ISWALTHERNOW 16
Seq2: 1 HEWASHEREBUTNOWISHERE 21

Local

Seq1: 1 W--HERE 5 Seq2: 3 WASHERE 9
Seq1: 1 W--HERE 5 Seq2: 15 WISHERE 21

Basic Local Alignment Search Tool

- **local** alignments; isolated regions of similarity
- fast and sensitive
- breaks the query sequence into “words”
- word hits to database sequences extended in both directions
Nucleotide Words

Make a lookup table based on the word size.

11-mer
ATGCTGCTAGTCGATGACGTAGCTA
ATGCTGCTAGT
TGCTGCTAGTC
GCTGCTAGTGC
...

Protein Words

AIEKCYTGCTLAQEADDTA
AIE
IEK LEK, IDK, IQK, IER, IDR, etc
EKC
KCY Neighborhood words
CYT
...

Lookup table, including neighborhood words, is based on word size, score matrix, and threshold.
### Scoring Systems - Proteins (BLOSUM62)

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

The Scoring Systems Scoring Systems -- Proteins (BLOSUM62)

IEK: keep LEK? (threshold = 11)

IEK = 14

LEK = 12

### Word Hits & Extensions

**Nucleotide:** one exact match

\[ \text{GCTGCTAGTCG} \]

**Protein:** two matches within 40 residues

\[ \text{IEKCYTGCTLAQEADDTA} \]

\[ \text{IDKE A D} \]
Search Basics

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

BLAST Programs

What is your goal?

<table>
<thead>
<tr>
<th>Program</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>blastn</td>
<td>nucleotide X nucleotide</td>
</tr>
<tr>
<td>blastp</td>
<td>protein X protein</td>
</tr>
<tr>
<td>blastx</td>
<td>nucleotide X protein</td>
</tr>
<tr>
<td>tblastx</td>
<td>nucleotide X nucleotide</td>
</tr>
<tr>
<td>tblastn</td>
<td>protein X nucleotide</td>
</tr>
</tbody>
</table>

6 frame, translated nucleotide searches
More BLAST Programs

MegaBLAST
- batch nucleotide queries
- very similar sequences

Discontiguous MegaBLAST
- batch nucleotide queries
- divergent sequences

BLAST 2 Seq. (pairwise)

Templates for Discontiguous Words

<table>
<thead>
<tr>
<th>W</th>
<th>t</th>
<th>coding:</th>
<th>non-coding:</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>16</td>
<td>11011011011011101101111</td>
<td>11100101101101101101111</td>
</tr>
<tr>
<td>11</td>
<td>16</td>
<td>111101101101111011011101</td>
<td>11101101101101101101110111</td>
</tr>
<tr>
<td>12</td>
<td>16</td>
<td>1011011011011011011011101</td>
<td>111001011011001101101111</td>
</tr>
<tr>
<td>11</td>
<td>18</td>
<td>1011011011011011011011101</td>
<td>111001011011001101101111</td>
</tr>
<tr>
<td>12</td>
<td>18</td>
<td>1110100101100100100101101</td>
<td>1001011001011100101101101</td>
</tr>
<tr>
<td>11</td>
<td>21</td>
<td>1110100101100100100101101</td>
<td>1110100101100100100101101</td>
</tr>
<tr>
<td>12</td>
<td>21</td>
<td>1001011011011100100101101</td>
<td>1110100101100100100101101</td>
</tr>
</tbody>
</table>

W = word size; # matches in template
W = template length

Search Basics

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

BLAST Databases – Nucleotide

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>nr</td>
<td>GenBank+EMBL+DDBJ +PDB sequences (no EST, STS, GSS, environmental samples or phase 0, 1 or 2 HTGS)</td>
</tr>
<tr>
<td>refseq_rna</td>
<td>RefSeq RNA sequences (NM_, XM_)</td>
</tr>
<tr>
<td>refseq_genomic</td>
<td>RefSeq genomic sequences (NC_, NT_, NG_)</td>
</tr>
<tr>
<td>month</td>
<td>new or revised GenBank+EMBL+DDBJ +PDB sequences from the last 30 days</td>
</tr>
<tr>
<td>est</td>
<td>Expressed Sequence Tags (EST division)</td>
</tr>
<tr>
<td>dbsts</td>
<td>Sequence Tag Site sequences (STS division)</td>
</tr>
<tr>
<td>htgs</td>
<td>Unfinished High Throughput Genomic Sequences, phases 0, 1 and 2 (finished phase 3 go to nr)</td>
</tr>
</tbody>
</table>
**HTGS - High Throughput Genomic Sequences**

<table>
<thead>
<tr>
<th>Phase 0</th>
<th>single-pass reads of a single clone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1</td>
<td>unfinished, unordered and contain gaps</td>
</tr>
<tr>
<td>Phase 2</td>
<td>unfinished, ordered and contain gaps</td>
</tr>
<tr>
<td>Phase 3</td>
<td>high quality finished sequences no gaps</td>
</tr>
</tbody>
</table>

**More BLAST Databases - Nucleotide**

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>gss</td>
<td>genome survey sequences</td>
</tr>
<tr>
<td>pdb</td>
<td>Protein Data Bank nucleotide sequences</td>
</tr>
<tr>
<td>chromosome</td>
<td>equivalent to refseq_genomic; complete genomes, chromosomes, contigs</td>
</tr>
</tbody>
</table>
# BLAST Databases – Protein

<table>
<thead>
<tr>
<th>nr</th>
<th>non-redundant GenBank CDS translations+PDB +SwissProt+PIR+PRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>refseq</td>
<td>RefSeq proteins</td>
</tr>
<tr>
<td>month</td>
<td>new or revised GenBank CDS translation+PDB +SwissProt+PIR+PRF released in the last 30 days</td>
</tr>
<tr>
<td>Swissprot</td>
<td>last major release of the SWISS-PROT database</td>
</tr>
<tr>
<td>pdb</td>
<td>Protein Data Bank protein sequences</td>
</tr>
<tr>
<td>env_nr</td>
<td>sequences from environmental samples</td>
</tr>
</tbody>
</table>

# BLAST Databases – Genome-specific

**Genomes**

- Human, mouse, rat, chimp, cow, pig, dog, sheep, cat
- Chicken, puffer fish, zebrafish
- Fly, honey bee, other insects
- Microbes, environmental samples
- Plants, nematodes
- Fungi, protozoa, other eukaryotes
Search Basics

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

Setting up a BLAST Search

Identifier or sequence
Your request has been successfully submitted and put into the Blast Queue.

Query = g[680251] (269 letters)
Hit the button to: **See conserved domains from CDD**

The request ID is: P177281941-31719-979-F371633-BLAST04

---

**BLAST Output**

BLASTP 2.2.15 [Oct-15-2006]

**Reference:**

**Reference:**

**Hit:** 1170083541-31719-9791471633-BLAST04

**Database:** non-redundant GenBank CDS
*translational*: ProteinPost-PIR-REF excluding environmental samples
4,500,144 sequences; 1,646,646,853 total letters

If you have any problems or questions with the results of this search please refer to the BLAST FAQ.

**Taxonomy reports**
**Tax BLAST Report**

- Lineage Report
- Genome Report
- Taxonomy Report

**Lineage Report**

- NCBI Minicourses
- Taxonomy Reports

**NCBI Minicourses**

**BLAST Output - Graphical Overview**

**Distribution of 514 Blast Hits on the Query Sequence**

P09079 Homeobox protein Hox-B5 (Hox-2.1) (MU-1) (P.Q. 1-240) E=119 E=5.8e-116
### BLAST Output - Descriptions

**Bit Score**

- derived from raw alignment score
- normalized, so can be used to compare different searches

**Expect value (E)**

- number of alignments expected to occur by chance

\[
E = mn^{-\frac{S}{2}}
\]

**Sample BLAST Output**

<table>
<thead>
<tr>
<th>Query</th>
<th>Subject</th>
<th>Bit Score</th>
<th>Expect Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>6680151</td>
<td>ref</td>
<td>NP_032924.1</td>
</tr>
<tr>
<td>gi</td>
<td>62650904</td>
<td>ref</td>
<td>NP_97184.1</td>
</tr>
<tr>
<td>gi</td>
<td>73909211</td>
<td>ref</td>
<td>NP_56178.1</td>
</tr>
<tr>
<td>gi</td>
<td>82504198</td>
<td>ref</td>
<td>NP_025138.1</td>
</tr>
<tr>
<td>gi</td>
<td>102887101</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>70770806</td>
<td>ref</td>
<td>NP_0102506.1</td>
</tr>
<tr>
<td>gi</td>
<td>145635107</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>128557916</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>5548533</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>75232973</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>145635107</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**E is dependent on search space**
BLAST Output - Alignments

>gi|1706354|sp|P09079|HXB5_MOUSE  Homeobox protein Hox-B5 (Hox-2.1) (MUS-Man) gi|6880251|ref|NP_032294.1|  Hox box B5 [Mus musculus] gi|522395|sp|AAA37842.1|  Homeobox protein
Length=269

Score = 419 bits (1077), Expect = 6e-116, Identities = 269/269 (100%), Positives = 269/269 (100%), Gaps = 0/269 (0%)

Query 1  MSSYVFNSFSGRYPNGFDqllnygsqssllspsyqLDPAAMHTGSGYNYNMDLSVRSS  60  
      MSSYVFNSFSGRYPNGFDyqllyngqssslspsyqLDPAAMHTGSGYNYNMDLSVRSS

Subjct 1  MSSYVFNSFSGRYPNGFDqllnygsqssllspsyqLDPAAMHTGSGYNYNMDLSVRSS  60  
      MSSYVFNSFSGRYPNGFDyqllyngqssslspsyqLDPAAMHTGSGYNYNMDLSVRSS

Query 61  AShHEGAUSESSAFAAPAFFKFPFCAcsscqspslpslpcpGDSKGAAPSASSPDQ  120  
      AShHEGAUSESSAFAAPAFFKFPFCAcsscqspslpslpcpGDSKGAAPSASSPDQ

Subjct 61  AShHEGAUSESSAFAAPAFFKFPFCAcsscqspslpslpcpGDSKGAAPSASSPDQ  120  
      AShHEGAUSESSAFAAPAFFKFPFCAcsscqspslpslpcpGDSKGAAPSASSPDQ

Query 121  ATPASSANFTE1Ehassassepaasq1spslaraspEMATSTAAEPGTIPFWM  180  
      ATPASSANFTE1Ehassassepaasq1spslaraspEMATSTAAEPGTIPFWM

Subjct 121  ATPASSANFTE1Ehassassepaasq1spslaraspEMATSTAAEPGTIPFWM  180  
      ATPASSANFTE1Ehassassepaasq1spslaraspEMATSTAAEPGTIPFWM

positives, from score matrix

BLAST Output - Alignments

>gi|62526079|dbj|BAD95556.1|  Hoxb-5 [Gallus gallus]  
Length=257

Score = 294 bits (753), Expect = 2e-78, Method: Composition-based stats. Identities = 204/262 (77%), Positives = 221/262 (84%), Gaps = 5/262 (2%)

Query 1  MSSYVFNSFSGRYPNGDFyqllyngsqsllspsyqRDFPAAMHTGSGYNYNMDLSVRSS  60  
      MSSYVFNSFSGRYPNGDFyqllyngsqsllspsyqRDFPAAMHTGSGYNYNMDLSVRSS

Subjct 1  MSSYVFNSFSGRYPNGDFyqllyngsqsllspsyqRDFPAAMHTGSGYNYNMDLSVRSS  60  
      MSSYVFNSFSGRYPNGDFyqllyngsqsllspsyqRDFPAAMHTGSGYNYNMDLSVRSS

low complexity
Search Basics

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

Advanced Options

Options for advanced blasting

Limit by
entrez query
Composition-
based
statistics
Choose filter
Low complexity
Expect
Word Size
Matrix
PSSM
Other advanced
PHI pattern

Example Entrez Query:
- Limit to Organism
- Nucleotides
- Other Advanced
- Expect value -e 10000
- descriptions -v 2000
- alignments -b 2000
Other Advanced Options

- e  expect value
- W  word size
- v  descriptions
- b  alignments

---

For short queries: increase e and lower W

Other advanced options:

- e 10000 – W 7 – v 2000 – b 1500

BLASTP –

“...short, nearly exact matches”

Options for advanced blasting

Limit by
entrez query

Compositional adjustments

Choose filter

- Low complexity
- Mask for lookup table only
- Mask lower case

Expect

Word Size

Matrix

PAM30
Scoring Matrices

Nucleotide search: identity matrix

\[
\begin{array}{cccc}
A & T & C & G \\
A & 1 & -3 & -3 & -3 \\
T & -3 & 1 & -3 & -3 \\
C & -3 & -3 & 1 & -3 \\
G & -3 & -3 & -3 & 1 \\
\end{array}
\]

Scoring Matrices

Protein search: substitution matrix

- Percent Accepted Mutation (PAM)
- Blocks Substitution Matrix (BLOSUM)
PAM Matrices

- global alignments of closely related proteins
- PAM1 calculated from sequences with no more than 1% divergence
- Other PAM matrices are extrapolated from PAM1
- Examples: PAM30 and PAM70

BLOSUM Matrices

- local alignments
- all matrices based on observed alignments; not extrapolated
- BLOSUM 62 calculated from sequences with no more than 62% identity
- Examples: BLOSUM45, BLOSUM62 and BLOSUM80
- BLOSUM62: very good for detecting weak protein similarities
- BLOSUM62 is the default matrix in BLAST
Scoring Systems - Proteins (BLOSUM62)

Substitution Matrices
A Provisional Table

<table>
<thead>
<tr>
<th>Query Length</th>
<th>Substitution Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;35</td>
<td>PAM30</td>
</tr>
<tr>
<td>35-50</td>
<td>PAM70</td>
</tr>
<tr>
<td>50-85</td>
<td>BLOSUM80</td>
</tr>
<tr>
<td>&gt;85</td>
<td>BLOSUM62</td>
</tr>
</tbody>
</table>
Search Basics

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples

Formatting Options

Format

Show
- Graphical Overview
- Linkout
- Sequence Retrieval
- NCBI-sq
- Alignment
- in HTML
- format

CDS feature

Masking Character
- Lower Case
- Masking Color
- Grey

Number of:
- Descriptions: 500
- Alignments: 250

Alignment view

Format for PSI-BLAST

Limit results by entrez query

Pairwise
- Pairwise with identities
- query-anchored with identities
- query-anchored without identities
- flat query-anchored with identities
- flat query-anchored without identities

Hit Table
### Flat Query-anchored with Identities

<table>
<thead>
<tr>
<th>Query</th>
<th>Identities</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSSYFNSF-8GRYPN-GDYqllqgss1gsy-RGS----AAMH----TGSY-----G</td>
<td>46</td>
</tr>
<tr>
<td>74225740</td>
<td>46</td>
</tr>
<tr>
<td>84571915</td>
<td>46</td>
</tr>
<tr>
<td>1138933</td>
<td>46</td>
</tr>
<tr>
<td>12842720</td>
<td>7</td>
</tr>
<tr>
<td>7644496</td>
<td>46</td>
</tr>
<tr>
<td>7723540</td>
<td>46</td>
</tr>
<tr>
<td>26526879</td>
<td>46</td>
</tr>
<tr>
<td>26526879</td>
<td>46</td>
</tr>
<tr>
<td>26526879</td>
<td>46</td>
</tr>
<tr>
<td>26526879</td>
<td>46</td>
</tr>
<tr>
<td>26526879</td>
<td>46</td>
</tr>
<tr>
<td>26526879</td>
<td>46</td>
</tr>
<tr>
<td>26526879</td>
<td>46</td>
</tr>
<tr>
<td>34267149</td>
<td>47</td>
</tr>
<tr>
<td>36737</td>
<td>47</td>
</tr>
<tr>
<td>53741666</td>
<td>47</td>
</tr>
<tr>
<td>25316136</td>
<td>46</td>
</tr>
<tr>
<td>4322074</td>
<td>46</td>
</tr>
<tr>
<td>7426738</td>
<td>46</td>
</tr>
<tr>
<td>1489444</td>
<td>46</td>
</tr>
<tr>
<td>5052652</td>
<td>46</td>
</tr>
<tr>
<td>3287931</td>
<td>46</td>
</tr>
<tr>
<td>4267926</td>
<td>46</td>
</tr>
</tbody>
</table>

### HitTable

<table>
<thead>
<tr>
<th># Fields: query id, subject ids, % identity, % positives, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score</th>
</tr>
</thead>
<tbody>
<tr>
<td>6605261</td>
</tr>
<tr>
<td>6605261</td>
</tr>
<tr>
<td>6605261</td>
</tr>
<tr>
<td>6605261</td>
</tr>
<tr>
<td>6605261</td>
</tr>
<tr>
<td>6605261</td>
</tr>
</tbody>
</table>

# Fields: query id, subject ids, % identity, % positives, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
Retrieval Results for an RID

Special
- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobulin BLAST (IgBlast)
- SNP BLAST

Meta
- Retrieve results

NCBI Search Basics

- Programs
- Databases
- Submit a search
- Interpret the results
- BLAST options
- Format options
- Examples
Examples

• Protein searches **more sensitive** than nucleotide searches
  
  - redundancy of the genetic code allows the same information to be encoded differently
  
• blastn best when searching within same organism

---

**BLASTN vs BLASTP**

*BLAST 2 SEQUENCES RESULTS VERSION BLASTP 2.2.14 [Apr-09-2006]*

Matrix Editor

<table>
<thead>
<tr>
<th>gap open</th>
<th>gap extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>1</td>
</tr>
</tbody>
</table>

x_dropoff: 50  expect: 10.0000  wordsize: 3  Filter  View option: Standard

Masking character option: X for protein, n for nucleotide  Masking color option: Black

Show CDS translation: Align

BLASTP

**NOTE:** Bitscore and expect value are calculated based on the size of the nr database.

Score = 45.8 bits (114),  Expect = 7e-04
Ids = 59/181 (32%), Positives = 90/181 (49%), Gaps = 27/181 (15%)
BLASTN vs TBLASTX

*Anopheles gambiae* mitochondrion, complete genome.
ACCESSION NC_002084 GI:5834911

*Homo sapiens* mitochondrion, complete genome.
ACCESSION NC_001807 GI:17981852

Additional Questions?

blast-help@ncbi.nlm.nih.gov
Hands-On