

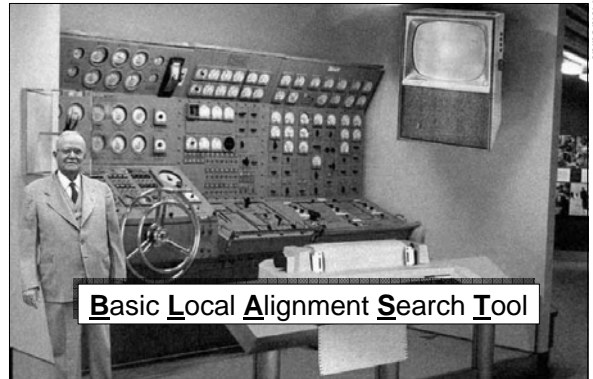
NCBI Molecular Biology Resources

Using NCBI BLAST

March 2007

Peter Cooper

Sequence Similarity Searching



Basic Local Alignment Search Tool

What BLAST tells you

- BLAST reports surprising alignments
 - Different than chance
- Assumptions
 - Random sequences
 - Constant composition
- Conclusions
 - Surprising similarities imply **evolutionary homology**

Evolutionary Homology: descent from a common ancestor
Does not always imply similar function

Basic Local Alignment Search Tool

- Widely used similarity search tool
- Heuristic approach based on Smith Waterman algorithm
- Finds best local alignments
- Provides statistical significance
- All combinations (DNA/Protein) query and database.
 - DNA vs DNA
 - DNA translation vs Protein
 - Protein vs Protein
 - Protein vs DNA translation
 - DNA translation vs DNA translation
- www, standalone, and network clients

Megablast: NCBI's Genome Annotator

NCBI Fieldguide

- Long alignments for similar DNA sequences
- Concatenation of query sequences
- Faster than blastn
- Contiguous Megablast
 - exact word match
 - Word size 28
- Discontiguous Megablast
 - initial word hit with mismatches
 - cross-species comparison

Templates for Discontiguous Words

NCBI Fieldguide

```

W = 11, t = 16, coding:      [1][1]0[1][1]0[1][1]0[1][1]0[1][1]
W = 11, t = 16, non-coding: 1110010110110111
W = 12, t = 16, coding:      1111101101101101
W = 12, t = 16, non-coding: 1110110110110111
W = 11, t = 18, coding:      101101100101101101
W = 11, t = 18, non-coding: 111010010110010111
W = 12, t = 18, coding:      101101101101101101
W = 12, t = 18, non-coding: 111010110010110111
W = 11, t = 21, coding:      100101100101100101101
W = 11, t = 21, non-coding: 111010010100010010111
W = 12, t = 21, coding:      100101101101100101101
W = 12, t = 21, non-coding: 111010010110010010111
    
```

W = word size; # matches in template
t = template length (window size within which the word match is evaluated)

Reference: Ma, B, Tromp, J, Li, M. PatternHunter: faster and more sensitive homology search. Bioinformatics March, 2002; 18(3):440-5

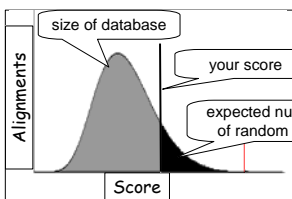
Local Alignment Statistics

NCBI Fieldguide

High scores of local alignments between two random sequences follow the Extreme Value Distribution

Expect Value

E = number of database hits you expect to find by chance



$$E = Kmne^{-\lambda S} \quad \text{or} \quad E = mn2^{-S'}$$

K = scale for search space
 λ = scale for scoring system
 S' = bitscore = $(\lambda S - \ln K) / \ln 2$

(applies to ungapped alignments)

Scoring Systems

NCBI Fieldguide

•Position Independent Matrices

•Nucleic Acids – identity matrix

•Proteins

- PAM Matrices (Percent Accepted Mutation)
 - Implicit model of evolution
 - Higher PAM number all calculated from PAM1
 - PAM250 widely used

•BLOSUM Matrices (BLOck SUBstitution Matrices)

- Empirically determined from alignment of conserved blocks
- Each includes information up to a certain level of identity
- BLOSUM62 widely used

•Position Specific Score Matrices (PSSMs)

- PSI and RPS BLAST

Scores

	V	D	S	-	C	Y	
	V	E	T	L	C	F	
BLOSUM62	+4	+2	+1	-12	+9	+3	<u>7</u>
PAM30	+7	+2	0	-10	+10	+2	<u>11</u>

NCBI FieldGuide

WWW BLAST

NCBI FieldGuide

The BLAST homepage

The screenshot shows the NCBI BLAST homepage with several sections highlighted:

- Nucleotide**:
 - Quickly search for highly similar sequences (megablast)
 - Quickly search for divergent sequences (discontiguous megablast)
 - Nucleotide-nucleotide BLAST (blastn)
 - Search for a sequence with its translated or translated sequence
- Protein**:
 - Protein-protein BLAST (blastp)
 - Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
 - Search for short, nearly exact matches
 - Search the conserved domain database (blastc)
- Standard databases**:
 - Human, mouse, rat, chimp, cow, dog, sheep, cat
 - Chicken, puffer fish, zebrafish
 - Environmental samples
 - Protoboa
 - Human, non-human, chimp, local, mouse
- Specialized Databases**:
 - Translated query vs. protein database (blastx)
 - Protein query vs. translated database (tblastn)
 - Translated query vs. translated database (tblastx)
- Special**:
 - Search for gene expression data (GEO BLAST)
 - High level sequences (BL2seq)
 - Screen for vector contamination (VecScreen)
 - Immunoglobulin BLAST (igblast)
 - SWI-BLAST
- Meta**:
 - Retrieve results

NCBI FieldGuide

BLAST Databases: Non-redundant protein

The screenshot shows the NCBI BLAST protein-protein search interface. A dropdown menu for "Choose database" is open, showing the following options:

- nr (non-redundant protein sequences)
 - GenBank CDS translations
 - NP_ RefSeqs
 - Outside Protein
 - PIR, Swiss-Prot, PRF
 - PDB (sequences from structures)
- pat protein patents
- env_nr environmental samples

NCBI FieldGuide

Nucleotide Databases: Genomic

Human and mouse genomes and reference transcripts now available

Human genomic plus transcript
 Mouse genomic plus transcript
 Others (nr etc.):
 Human genomic plus transcript

NEW Two new **Human** and **Mouse** databases combine genomic plus transcript alignments in a single report. You can also choose from **Others** to use nr or an existing database.

Now: **BLAST** or **FASTA** or **MEGALAN**

Nucleotide Databases: Traditional

Human and mouse genomes and reference transcripts now available

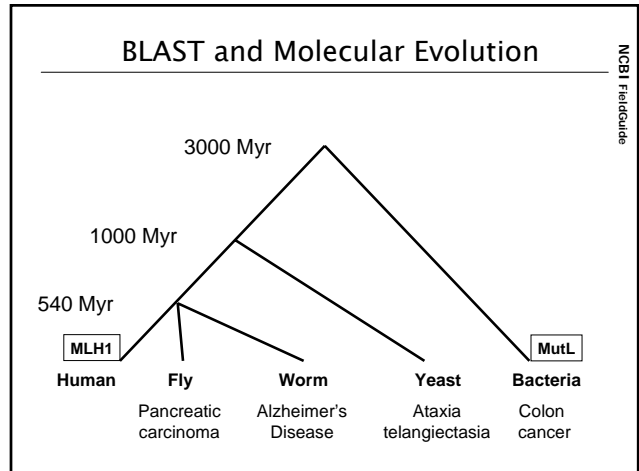
Human genomic plus Transcript
 Human genomic plus transcript
 Mouse genomic plus transcript
 Others (nr etc.):
 refseq_ma
 refseq_genomic
 est
 est_human
 est_mouse
 est_others
 gss
 htgs
 pat
 pab
 month
 alu_repeats
 dbsts
 chromosome
 wgs

NEW Two new **Human** and **Mouse** databases combine genomic plus transcript alignments in a single report. You can also choose from **Others** to use nr or an existing database.

Now: **BLAST** or **FASTA** or **MEGALAN**

Nucleotide Databases: Traditional

- **nr (nt)**
 - Traditional GenBank
 - NM_ and XM_ RefSeqs
 - **refseq_rna**
- **refseq_genomic**
 - NC_ RefSeqs
- **dbest**
 - EST Division
 - **est_human**, **mouse**, **others**
- **htgs**
 - HTG division
- **gss**
 - GSS division
- **wgs**
 - whole genome shotgun
- **env_nt**
 - environmental samples



Protein BLAST Page

NCBI **protein-protein BLAST**
Nucleotide Protein Translations Retrieve results for an RID

Search: **>Mutated in Colon Cancer**
 IETVYAAYLKPNTHFFLYLSLEISQNVVDVNVHPTKHEVHFLHEESILE
 VQQHIESKLLGSSSRMYFTQLLPGLAGPQSGEMVKSTTSLTSSTSGS
 DKVYAHQVVRTDSREQLDAFLQPLSKPLSS

Set subsequence From: _____ To: _____

Choose database: **swissprot** **Protein database**

Do CD-Search

Now: **BLAST** or **blastp** **blastx**

NCBI Fieldguide

Advanced Options: Entrez limit

Options for advanced blasting

Limit by entrez query: _____ AND _____

Composition-based statistics: **all[Filter] NOT mammals[Organism]**

gene_in_mitochondrion[Properties]
 2003:2005 [Modification Date]
 tpa[Filter]

Nucleotide
 biomol_mrna[Properties]
 biomol_genomic[Properties]

All organisms
 All organisms [ORGN]
 Viruses [ORGN]
 Archaea [ORGN]
 Bacteria [ORGN]
 Eukaryota [ORGN]
 Viridiplantae [ORGN]
 Fungi [ORGN]
 Metazoa [ORGN]
 Arthropoda [ORGN]
 Vertebrata [ORGN]
 Mammalia [ORGN]
 Rodents [ORGN]
 Primates [ORGN]
 E
 Aeropyrum pernix [ORGN]
 Aquifex aeolicus [ORGN]
 Arabidopsis thaliana [ORGN]
 Bacillus subtilis [ORGN]
 Bos taurus [ORGN]
 Caenorhabditis elegans [ORGN]

NCBI Fieldguide

Advanced Options: Filters

Protein

Choose filter Low complexity Mask for lookup table only Mask lower case

Hides low complexity for initial word hits only

Masks Low Complexity Sequence with X or n

Masks regions of query in lower case (pre-masked)

Nucleotide

Choose filter Low complexity Repeats: **Human** Mask for lookup table only Mask lower case

Masks Human or Mouse Interspersed repeats. Default for genome searches.

NCBI Fieldguide

Advanced Options: composition based stats

Options for advanced blasting

Limit by entrez query: _____ AND _____

Composition-based statistics: **Composition-based statistics**

Choose filter: **Composition-based statistics**

Expect: **10**

Word Size: **3**

Matrix: **BLOSUM62** Gap Costs: _____

Amino acid composition: **Histone H1**

Ala (A)	42	19.6%
Arg (R)	4	1.9%
Asn (N)	4	1.9%
Asp (D)	1	0.5%
Cys (C)	0	0.0%
Gln (Q)	2	0.9%
Glu (E)	6	2.8%
Gly (G)	13	6.1%
His (H)	0	0.0%
Ile (I)	3	1.4%
Leu (L)	10	4.7%
Lys (K)	57	26.6%
Met (M)	0	0.0%
Phe (F)	1	0.5%
Pro (P)	19	8.9%
Ser (S)	23	10.7%
Thr (T)	14	6.5%
Trp (W)	0	0.0%
Tyr (Y)	1	0.5%
Val (V)	14	6.5%

Negatively charged residues (Asp + Glu): 7
 Positively charged residues (Arg + Lys): 61

NCBI Fieldguide

BLAST Formatting Page

formatting **BLAST**

Nucleotide Protein Translations Retrieve results for an ID

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

Query = Mutated in Colon Cancer (131 letters)

Putative conserved domains have been detected, click on the image below for detailed results.

Conserved Domain

The request ID is 1030888657-012751-16006

[Format](#) or [rescore](#)

The results are estimated to be ready in 36 seconds but may be done sooner.

NCBI Fieldguide

BLAST Output: Graphical Overview

Distribution of 70 Blast Hits on the Query Sequence

NCBI Fieldguide

BLAST Output: Descriptions

Sorted by e values

Score E (bits) Value

Accession	Description	Score	E (bits)	Value
gi148474996 sp O99726 MLN1_SCHPO	Putative MUTL protein homo...	50	3e-10	0.0000000000
gi117090561 sp P201 MLN1_YEAST	MUTL protein	46	1.1e-08	0.0000000000
gi111710888 sp P4494 MUTL_HAEIN	DNA mismatch repair protei...	44	1e-08	0.0000000000
gi12484848 sp Q8RA70 MUTL_THETN	DNA mismatch repair protei...	43	1e-07	0.0000000000
gi13578866 sp P57886 MUTL_PASNU	DNA mismatch repair protei...	49	2e-06	0.0000000000
gi12097120 sp Q97120 MUTL_CLOAB	DNA mismatch repair protei...	47	6e-06	0.0000000000
gi18928224 sp P74925 MUTL_THENA	DNA mismatch repair protei...	47	1e-05	0.0000000000
gi132129762 sp Q87AC9 MUTL_XYLYT	DNA mismatch repair protei...	45	2e-05	0.0000000000
gi120139217 sp Q9KAC1 MUTL_BACHD	DNA mismatch repair protei...	44	5e-05	0.0000000000
gi137995551 sp Q87V21 MUTL_PSK3H	DNA mismatch repair protei...	44	7e-05	0.0000000000
gi120455148 sp Q9HUL8 MUTL_PSKAF	DNA mismatch repair protei...	44	7e-05	0.0000000000
gi120455107 sp Q82F86 MUTL_RICCN	DNA mismatch repair protei...	43	1e-04	0.0000000000
gi133301356 sp Q82119 MUTL_CHLCV	DNA mismatch repair protei...	43	1e-04	0.0000000000
gi120455160 sp Q9PFB8 MUTL_XYLFA	DNA mismatch repair protei...	42	2e-04	0.0000000000
gi125090732 sp Q8KX73 HUS1	DNA mismatch repair protei...	42	2e-04	0.0000000000
gi11278521 sp P23347 MUTL1	DNA mismatch repair protei...	42	3e-04	0.0000000000
gi120455084 sp Q82DM4 MUTL1	DNA mismatch repair protei...	42	3e-04	0.0000000000
gi129427778 sp Q8FAR9 MUTL_ECOL6	DNA mismatch repair protei...	42	3e-04	0.0000000000
gi120455147 sp Q9H8R6 MUTL_HALN1	DNA mismatch repair protei...	40	0.99	0.0000000000
gi13246727 sp P14541 PHD1_YEAST	DNA mismatch repair protei...	29	2.2	0.0000000000
gi1325231 sp P02139 LGB1_LUPLU	Leghemoglobin I	2.0	2.9	0.0000000000
gi156749250 sp Q6LV74 META_PHOPE	Homoerectine O-succinyltrac...	2.0	4.9	0.0000000000
gi1326230 sp P02240 LGB2_LUPLU	Leghemoglobin II	2.7	6.4	0.0000000000
gi134285971 sp O01090 T920_TREPA	Hypothetical protein TP0920	2.7	0.4	0.0000000000

NCBI Fieldguide

TaxBLAST: Taxonomy Reports

<i>Chlamydomonas reinhardtii</i> [Chlamydomonadophyceae] taxid: 89557	42	1e-04
gi133301356 sp Q82119 MUTL_CHLCV DNA mismatch repair prote...	42	1e-04
<i>Spizella monticola</i> [Spizellidae] taxid: 2371	41	1e-04
gi120455107 sp Q82F86 MUTL_RICCN DNA mismatch repair prote...	41	1e-04
<i>Chlorobacterium tepidum</i> [Chlorobacteriales] taxid: 1097	41	1e-04
gi12097120 sp Q97120 MUTL_CLOAB DNA mismatch repair prote...	41	1e-04
<i>Escherichia coli</i> [Enterobacteriales] taxid: 562	41	1e-04
gi11278521 sp P23347 MUTL1 DNA mismatch repair protei...	41	1e-04
<i>Escherichia coli</i> [Enterobacteriales] taxid: 8334	41	1e-04
gi120455084 sp Q82DM4 MUTL1 DNA mismatch repair prote...	41	1e-04
<i>Escherichia coli</i> [Enterobacteriales] taxid: 217992	41	1e-04
gi129427778 sp Q8FAR9 MUTL_ECOL6 DNA mismatch repair prote...	41	1e-04
<i>Neisseria meningitidis serogroup B</i> [Neisseriaceae] taxid: 491	41	1e-04
gi120455107 sp Q82F86 MUTL_RICCN DNA mismatch repair prote...	41	1e-04
<i>Haemophilus parvulus</i> [Haemophilales] taxid: 1801	41	1e-04
gi120455107 sp Q82F86 MUTL_RICCN DNA mismatch repair prote...	41	1e-04
<i>Sinorhizobium meliloti</i> [Sinorhizobiales] taxid: 302	41	1e-04
gi120455107 sp Q82F86 MUTL_RICCN DNA mismatch repair prote...	41	1e-04
<i>Salmonella typhimurium</i> [Enterobacteriales] taxid: 602	40	6e-04
gi11278521 sp P23347 MUTL1 DNA mismatch repair protei...	40	6e-04
<i>Salmonella typhi</i> [Enterobacteriales] taxid: 601	40	6e-04
gi120455099 sp Q82107 MUTL_BALT1 DNA mismatch repair prote...	40	6e-04
<i>Xanthomonas axonopodis</i> pv. <i>citri</i> [Xanthomonadales] taxid: 92829	40	6e-04
gi120455107 sp Q82F86 MUTL_RICCN DNA mismatch repair prote...	40	6e-04
<i>Xanthomonas campestris</i> pv. <i>campestris</i> [Xanthomonadales] taxid: 340	40	6e-04
gi112097120 sp Q97120 MUTL_CLOAB DNA mismatch repair prote...	40	6e-04

NCBI Fieldguide

BLAST Output: Alignments

NCBI Fieldguide

```

>gi|127552|sp|P23367|MUTL_ECOLI DNA mismatch repair protein mutL
Length = 615

Score = 42.0 bits (97), Expect = 3e-04
Identities = 26/59 (44%), Positives = 33/59 (55%), Gaps = 9/59 (15%)

Query 9  LPKNTHPFLYLSLEISPNQVNVVHPTKHEVHF-----LHE--ESILEV-QQHIESKL 58
          L + P L LEI P VDVNVHF KHEV F +H+ + +L V QQ +E+ L
Sbjct 280 LGADQQPAFVLYLEIDPHQVDVNVHFAKHEVRFHQSRVLVHDFYQQGVLSVLQQLETPL 338
  
```

Low Complexity Filter

NCBI Fieldguide

```

>gi|730028|sp|P40692|MLH1_HUMAN DNA mismatch repair protein Mlh1
Length=756

Score = 231 bits (589), Expect = 1e-62
Identities = 131/131 (100%), Positives = 131/131 (100%), Gaps = 0/131 (0%)

Query 1  IETVYAALPKNTHPFLYLSLEISPNQVNVVHPTKHEVHFLEHESILERVQQHIESKLL 60
          IETVYAALPKNTHPFLYLSLEISPNQVNVVHPTKHEVHFLEHESILERVQQHIESKLL
Sbjct 276 IETVYAALPKNTHPFLYLSLEISPNQVNVVHPTKHEVHFLEHESILERVQQHIESKLL 335

Query 61  GNSSSRMVFTQTLPLGLAGPSGEMV*sttltststsgsskKVYAHQMVRTDSREQKLD 120
          GNSSSRMVFTQTLPLGLAGPSGEMV*sttltststsgsskKVYAHQMVRTDSREQKLD
Sbjct 336 GNSSSRMVFTQTLPLGLAGPSGEMV*sttltststsgsskKVYAHQMVRTDSREQKLD 395

Query 121 FLQPLSKPLSS 131 low complexity sequence filtered
          FLQPLSKPLSS
Sbjct 396 FLQPLSKPLSS 406
  
```

Nucleotide: Human Repeats

NCBI Fieldguide

Distribution of 20987 Blast Hits on the Query Sequence

Mouse-over to show define and scores, click to show alignments

Color key for alignment scores

Query <40 40-50 50-80 80-200 >=200

Accession	Species	Score (Bits)	E Value
gnl alu RS014568	Human Alu-3b subfamily consensus	454	2e-127
gnl alu RS014569	Human Alu-3b subfamily consensus	431	2e-120
gnl alu RS014572	Human Alu-3p subfamily consensus	427	3e-119
gnl alu RS014574	Human Alu-3x subfamily consensus	421	1e-117
gnl alu RS014571	Human Alu-3c subfamily consensus	404	2e-112
gnl alu RS014573	Human Alu-3g subfamily consensus	404	2e-112
gnl alu RS014570	Human Alu-3h subfamily consensus	392	7e-109
gnl alu RS014567	Human Alu-3 subfamily consensus	344	3e-100

Nucleotide: Human Repeat Filter

NCBI Fieldguide

Distribution of 20987 Blast Hits on the Query Sequence

Mouse-over to show define and scores, click to show alignments

Color key for alignment scores

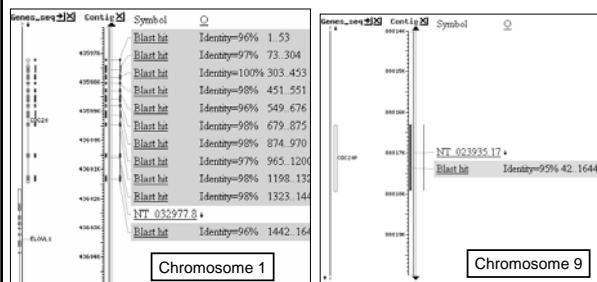
Choose filter Low complexity Repeats Human Mask for lookup table only Mask lower case

Query <40 40-50 50-80 80-200 >=200

AY728024 Homo sapiens setum albumin precursor, mRNA, complete cd. S=408 E=1.2e-11

Links to Map Viewer

NCBI FieldGuide



Genomic BLAST pages

Higher Genomes

NCBI FieldGuide

Service Addresses

NCBI FieldGuide

- **General Help** info@ncbi.nlm.nih.gov
- **BLAST** blast-help@ncbi.nlm.nih.gov

Telephone support: 301- 496- 2475