NCBI FieldGuide

NCBI Molecular Biology Resources

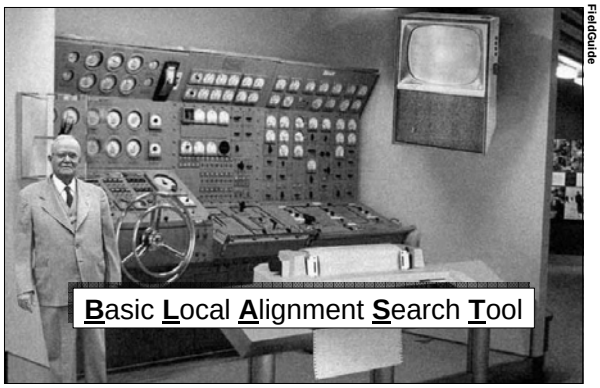
Using NCBI BLAST

March 2007

Peter Cooper

Sequence Similarity Searching

NCBI FieldGuide



What BLAST tells you

NCBI FieldGuide

- 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

NCBI FieldGuide

- 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

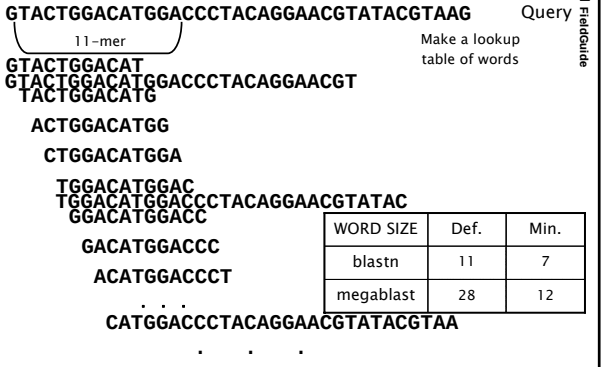
BLAST and BLAST-like programs

NCBI FieldGuide

- Traditional BLAST (blastall) nucleotide, protein, translations
  - blastn nucleotide query vs. nucleotide database
  - blastp protein query vs. protein database
  - blastx nucleotide query vs. protein database
  - tblastn protein query vs. translated nucleotide database
  - tblastx translated query vs. translated database
- Megablast nucleotide only
  - Contiguous megablast
    - **Nearly identical sequences**
  - Discontiguous megablast
    - **Cross-species comparison**
- Position Specific BLAST Programs protein only
  - Position Specific Iterative BLAST (PSI-BLAST)
    - **Automatically generates a position specific score matrix (PSSM)**
  - Reverse PSI-BLAST (RPS-BLAST)
    - **Searches a database of PSI-BLAST PSSMs**

Nucleotide Words

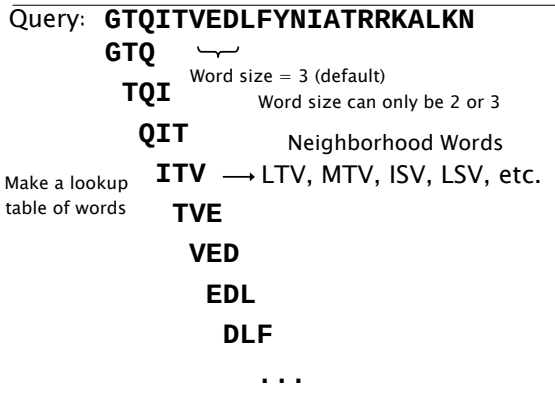
NCBI FieldGuide



WORD SIZE	Def.	Min.
blastn	11	7
megablast	28	12

Protein Words

NCBI FieldGuide



An alignment that BLAST can't find

NCBI FieldGuide

```
 1 GAATATATGAAGACCAAGATTGCAGTCCTGCTGGCCTGAACCACGCTATTCTTGCTGTTG
   || | || || || |  || || ||   ||  |  ||| |||||| | | || | ||| |
 1 GAGTGTACGATGAGCCCGAGTGTAGCAGTGAAGATCTGGACCACGGTGTACTCGTTGTCG

61 GTTACGGAACCGAGAATGGTAAAGACTACTGGATCATTAAGAACTCCTGGGGAGCCAGTT
   | || ||     ||  ||| ||  | |||||| || | ||||||  |||||  |     |
61 GCTATGGTGTTAAGGGTGGGAAGAAGTACTGGCTCGTCAAGAACAGCTGGGCTGAATCCT

121 GGGGTGAACAAGGTTATTTCAGGCTTGCTCGTGGTAAAAAC
    |||| || ||||| ||  ||    |  | ||||  || |||
121 GGGGAGACCAAGGCTACATCCTTATGTCCCGTGACAACAAC
```

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:	1101101101101101
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

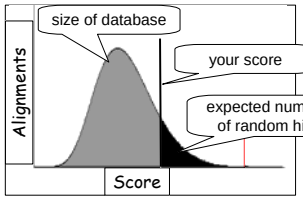
Local Alignment Statistics

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



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

K = scale for search space
λ = scale for scoring system
S' = bitscore = (λS - lnK)/ln2

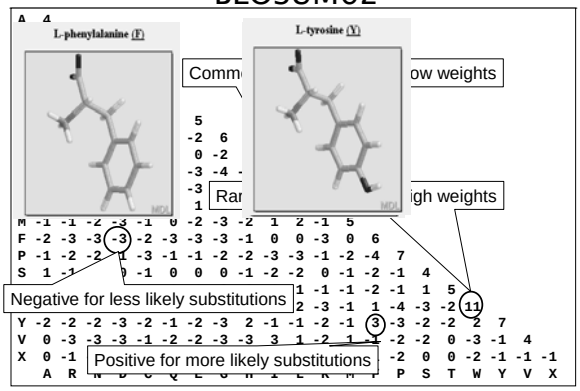
(applies to ungapped alignments)

Scoring Systems

NCBI FieldGuide

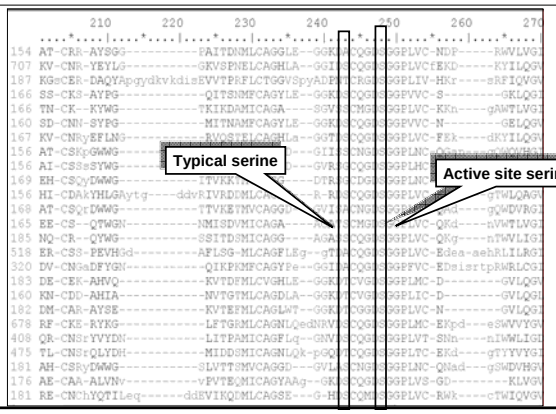
- **Position Independent Matrices**
  - **Nucleic Acids – identity matrix**
  - **Proteins**
    - **PAM Matrices (Percent Accepted Mutation)**
      - **Implicit model of evolution**
      - **Higher PAM number all calculated from PAM1**
      - **PAM250 widely used**
    - **BLOSUM Matrices (BLOck 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**

BLOSUM62



NCBI FieldGuide

Position Specific Substitution Rates



NCBI FieldGuide

Position Specific Score Matrix (PSSM)

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
206	D	0	-2	0	2	-4	2	4	-4	-3	-5	-4	0	-2	-6	1	0	-1	-6	-4	-1
207	G	-2	-1	0	-2	-4	-3	-3	6	-4	-5	-5	0	-2	-3	-2	2	-1	0	-6	-5
208	V	-1	1	-3	-3	-5	-1	-2	6	-1	-4	-5	1	-5	-6	-4	0	-2	-6	-4	-2
209	I	-3	3	-3	-4	-6	0	-1	-4	-1	2	-4	6	-2	-5	-5	3	0	-1	-4	0
210	S	-2	-5	0	8	-5	-3	-2	-1	-4	-7	-6	-4	-6	-7	-5	1	-3	-7	-5	-6
211	S	4	-4	-4	-4	-4	-1	-4	-2	-3	-3	-5	-4	-4	-5	-1	4	3	-6	-5	-3
212	C	-4	-7	-6	-7	12	-7	-7	-5	-6	-5	-5	-7	-5	0	-7	4	-4	-5	0	-4
213	N	-2	0	2	-1	-6	7	[illegible]								-5	1	-3	-3	-4	-3
214	G	-2	-3	-3	-4	-4	-4	[illegible]								-6	3	-5	-6	-6	-6
215	D	-5	-5	-2	9	-7	-4	[illegible]								-5	4	-4	-8	-7	-7
216	S	-2	-4	-2	-4	-4	-3	-3	-3	-4	-6	-6	-3	-5	-6	-4	7	-2	-6	-5	-5
217	G	-3	-6	-4	-5	-6	-5	-6	8	-6	-8	-7	-5	-6	-7	-6	4	-5	-6	-7	-7
218	G	-3	-6	[illegible]								7	-7	-5	-6	-7	-6	2	-4	-6	-7	-7
219	P	-2	-6	[illegible]								6	-7	-4	-6	-7	9	4	-4	-7	-7	-6
220	L	-4	-6	-7	-7	-5	-5	-6	-7	0	-1	6	-6	1	0	-6	6	-5	-5	-4	0
221	N	-1	-6	0	-6	-4	-4	-6	-6	-1	3	0	-5	4	-3	-6	2	-1	-6	-1	6
222	C	0	-4	-5	-5	10	-2	-5	-5	1	-1	-1	-5	0	-1	-4	1	0	-5	0	0
223	Q	0	1	4	2	-5	2	0	0	0	-4	-2	1	0	0	0	1	-1	-3	-3	-4
224	A	-1	-1	1	3	-4	-1	1	4	-3	-4	-3	-1	-2	-2	-3	0	-2	-2	-2	-3

Serine scored differently in these two positions

Active site nucleophile

NCBI FieldGuide

Gapped Alignments

- Gapping provides more biologically realistic alignments
- Gapped BLAST parameters must be simulated
- Affine gap costs = -(a+bk)
  - a = gap open penalty b = gap extend penalty
  - A gap of length 1 receives the score –(a+b)

NCBI FieldGuide

Scores

NCBI FieldGuide

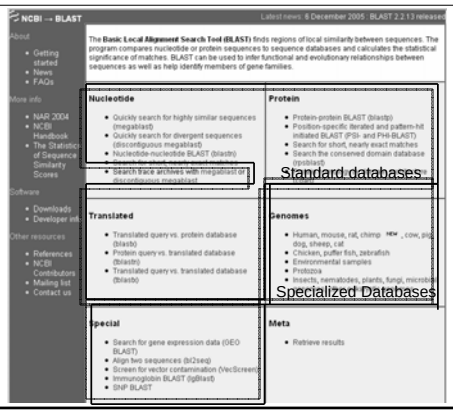
	V	D	S	–	C	Y	
	V	E	T	L	C	F	
BLOSUM62	+4	+2	+1	-12	+9	+3	**7**
PAM30	+7	+2	0	-10	+10	+2	**11**

NCBI FieldGuide

WWW BLAST

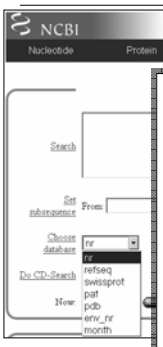
The BLAST homepage

NCBI FieldGuide



BLAST Databases: Non-redundant protein

NCBI FieldGuide

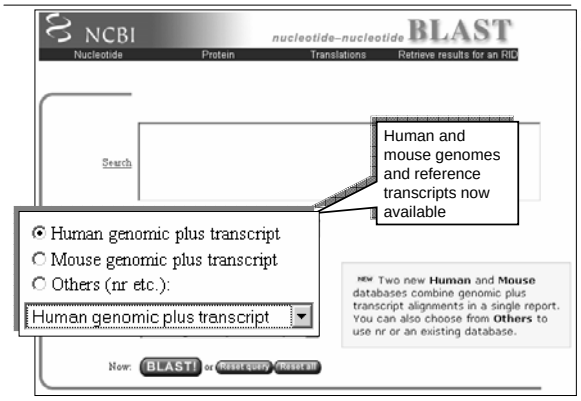


nr (non-redundant protein sequences)
- GenBank CDS translations
- NP_ **RefSeqs**
- Outside Protein
  - PIR, **Swiss-Prot**, PRF
  - **PDB** (sequences from structures)

pat protein patents

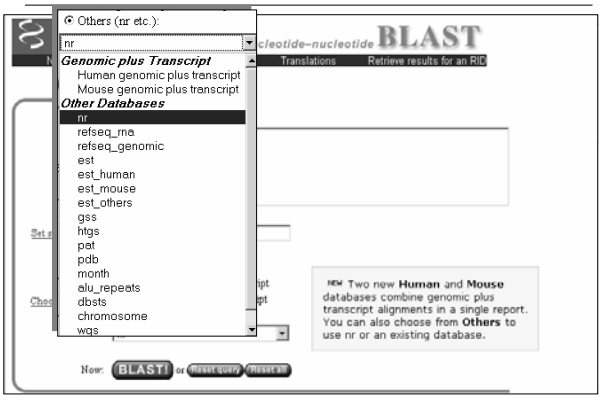
env_nr environmental samples

Nucleotide Databases: Genomic



NCBI FieldGuide

Nucleotide Databases: Traditional



NCBI FieldGuide

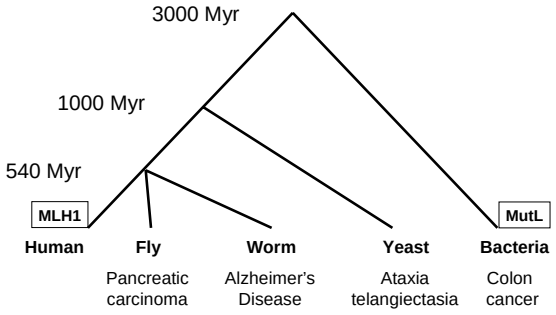
Nucleotide Databases: Traditional

- **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

NCBI FieldGuide

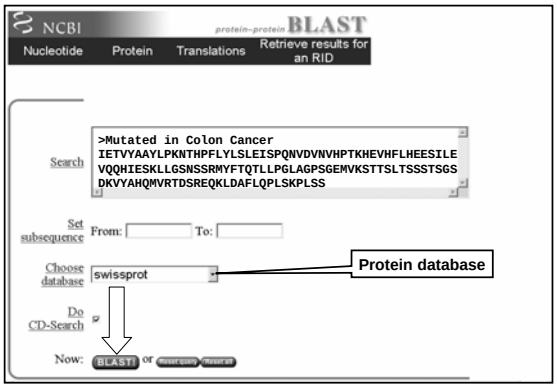
BLAST and Molecular Evolution



NCBI FieldGuide

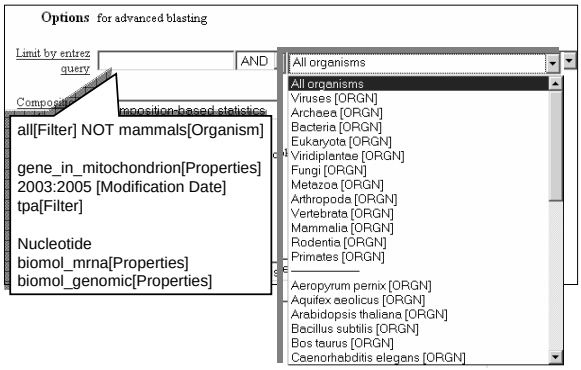
Protein BLAST Page

NCBI FieldGuide



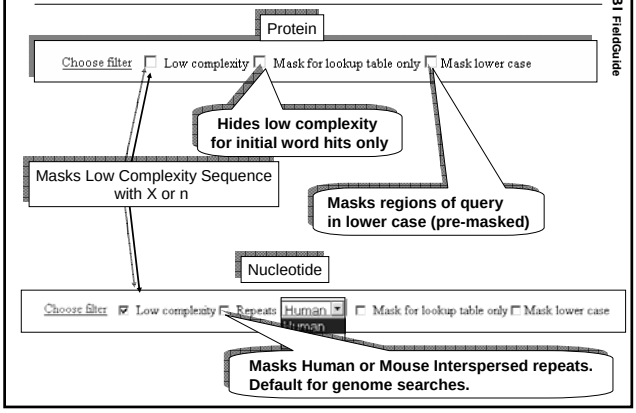
Advanced Options: Entrez limit

NCBI FieldGuide



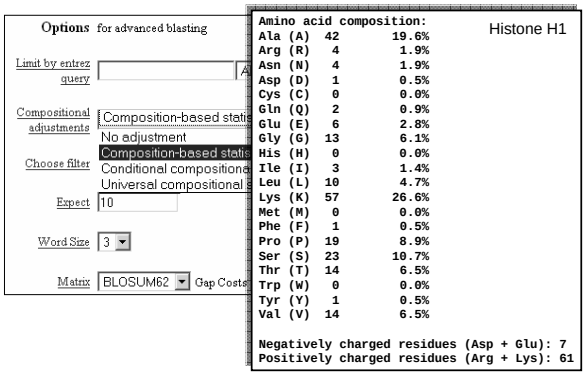
Advanced Options: Filters

NCBI FieldGuide

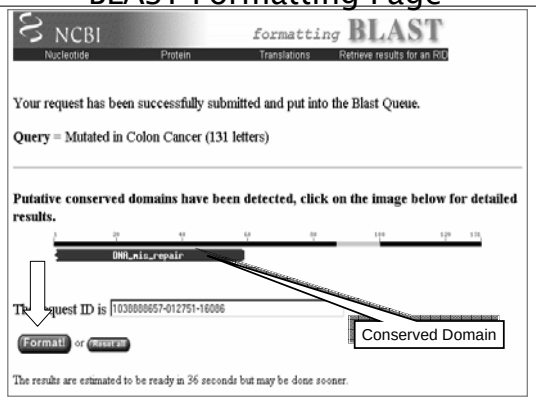


Advanced Options: Composition based stats

NCBI FieldGuide

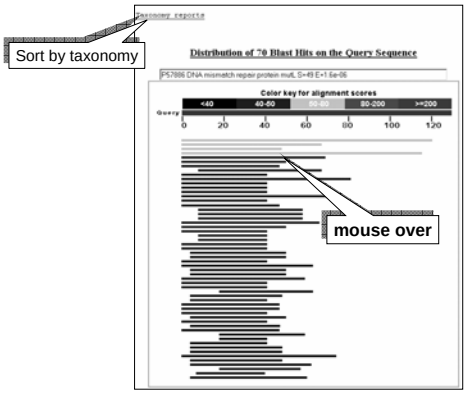


BLAST Formatting Page



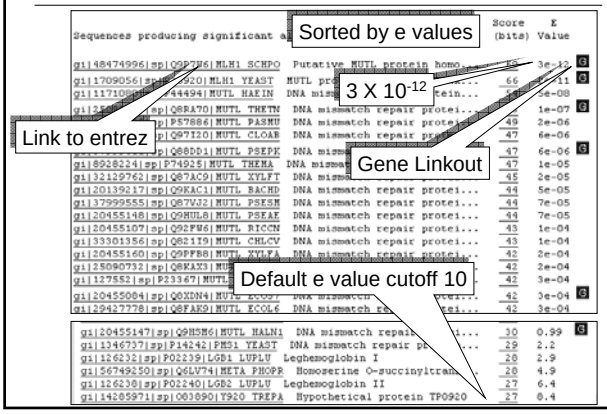
NCBI FieldGuide

BLAST Output: Graphical Overview



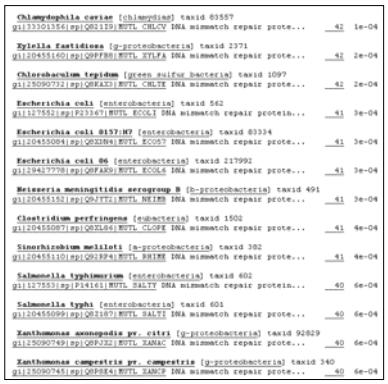
NCBI FieldGuide

BLAST Output: Descriptions



NCBI FieldGuide

TaxBLAST: Taxonomy Reports



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    LPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHF-----LHE---ESILEV-QQHIESKL  58
            L  +  P   L LEI P  VDVNVHP KHEV F     +H+    + +L V QQ +E+ L
Sbjct  280  LGADQQPAFVLYLEIDPHQVDVNVHPAKHEVRFHQSRLVHDFIYQGVLSVLQQQLETPL  338
```

Identical match

positive score (conservative)

negative substitution

gap

Low Complexity Filter

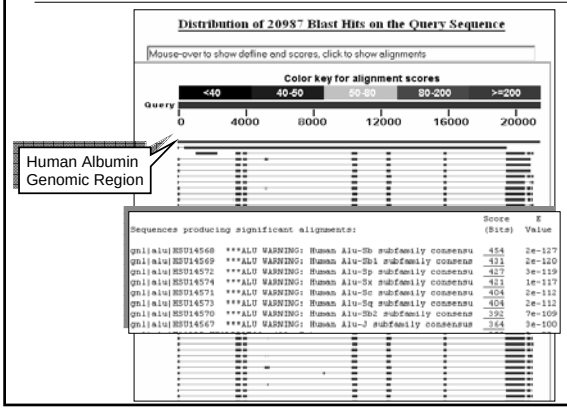
NCBI FieldGuide

```
>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  IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL  60
            IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL
Sbjct  276  IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL  335

Query   61  GSNSSRMYFTQTLLPGLAGPSGEMVKsttsltssstsgssDKVYAHQMVRTDSREQKLDA  120
            GSNSSRMYFTQTLLPGLAGPSGEMVKSTTSLTSSSTSGSSDKVYAHQMVRTDSREQKLDA
Sbjct  336  GSNSSRMYFTQTLLPGLAGPSGEMVKSTTSLTSSSTSGSSDKVYAHQMVRTDSREQKLDA  395

Query  121  FLQPLSKPLSS  131
            FLQPLSKPLSS
Sbjct  396  FLQPLSKPLSS  406
```

low complexity sequence filtered

Nucleotide: Human Repeats

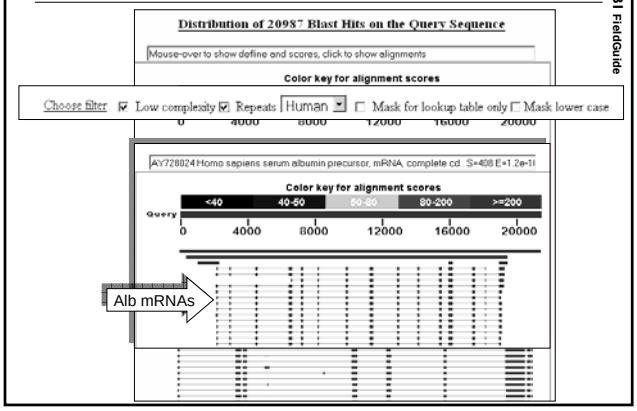
NCBI FieldGuide



Human Albumin Genomic Region

Nucleotide: Human Repeat Filter

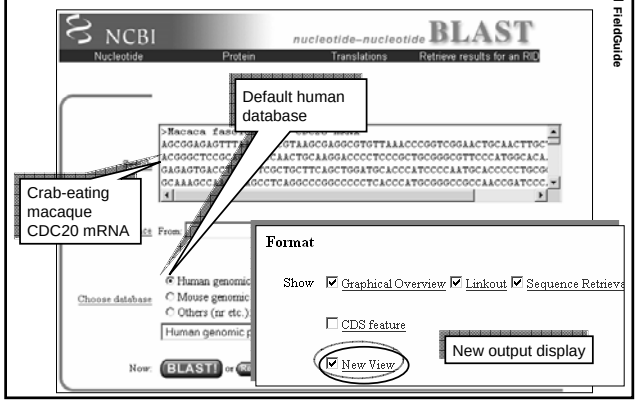
NCBI FieldGuide



Alb mRNAs

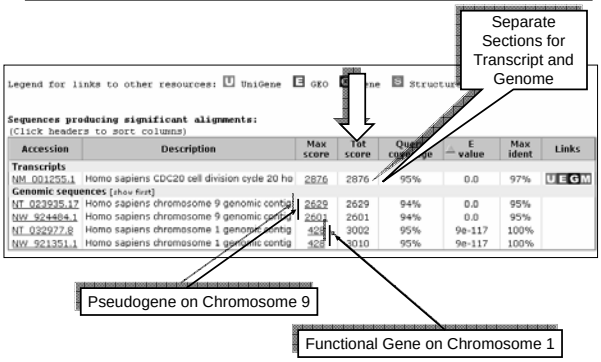
Nucleotide BLAST: New Output

NCBI FieldGuide



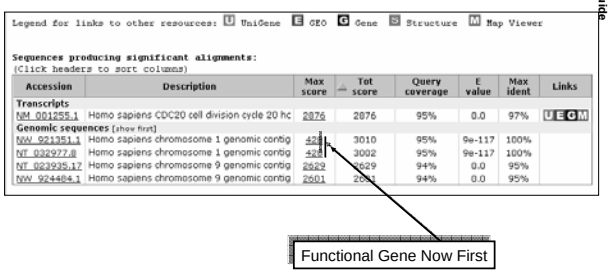
Sortable Results

NCBI FieldGuide



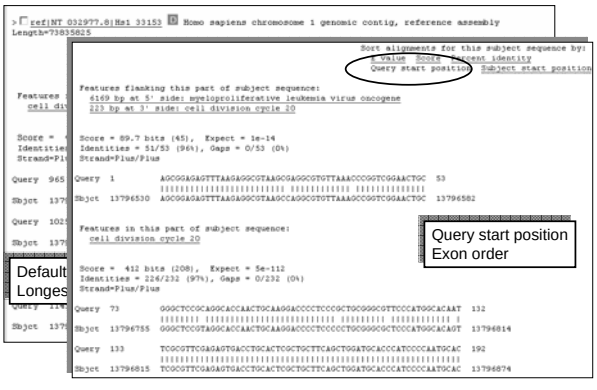
Total Score: All Segments

NCBI FieldGuide



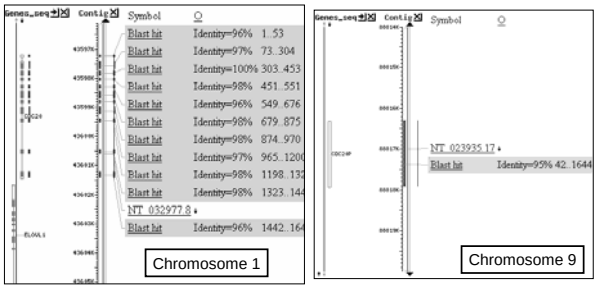
Sorting in Exon Order

NCBI FieldGuide



Links to Map Viewer

NCBI FieldGuide



NCBI FieldGuide

Genomic BLAST pages

Higher Genomes

Service Addresses

NCBI FieldGuide

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

Telephone support: 301- 496- 2475