

Searching Sequence Databases



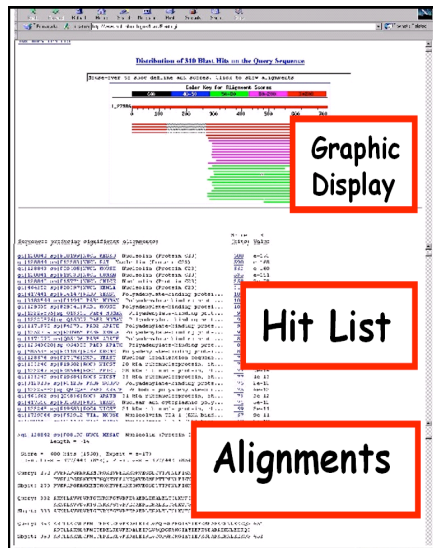
Excerpt modified by Dietlind Gerloff
for use in BME110/BIOL181
Winter 2008

© Wiley Publishing, 2007. All Rights Reserved.



Reading BLAST Output

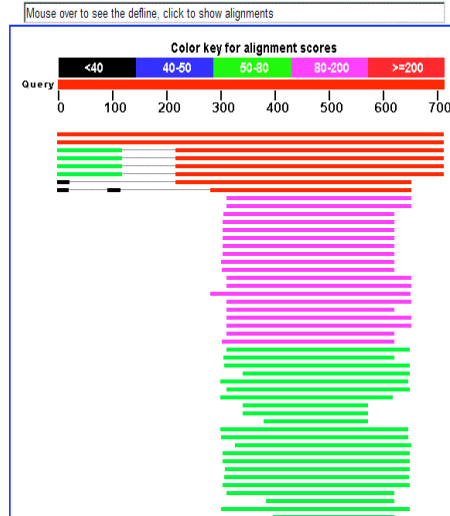
- ✓ Graphic Display
 - Overview of the alignments
- ✓ Hit List
 - Gives the score of each match
- ✓ Alignments
 - Details of each alignment



The Graphic Display

- ✓ The Horizontal Axis (0-700) corresponds to your protein (query)
- ✓ Color codes indicate that match's quality
 - Red: very good
 - Green: acceptable
 - Black: bad
- ✓ Thin lines join independent matches on the same sequence

Distribution of 297 Blast Hits on the Query Sequence



The Hit List

- ✓ Sequence accession number
 - Depends on the database
- ✓ Description
 - Taken from the database
- ✓ Bit score
 - High bit score = good match
- ✓ E-Value
 - Low E-value = good match
- ✓ Links
 - Genome
 - Uniref, database of transcripts

Distance tree of results [New](#) [Related Structures](#)

Sequences producing significant alignments:	Score (Bits)	E Value	
ref XP_516145.2 PREDICTED: hypothetical protein [Pan troglodyte	803	0.0	G
ref XP_001116949.1 PREDICTED: similar to nucleolin [Macaca mula	793	0.0	UG
sp Q4R437 NUCL_MACFA Nucleolin >dbj BAE00345.1 unnamed prote...	746	0.0	UG
ref NP_005372.2 nucleolin [Homo sapiens] >sp P19338 NUCL_HIM...	744	0.0	UG
sp Q5RF26 NUCL_POMFY Nucleolin >emb CAH84631.1 hypothetical ...	739	0.0	UG
gb AAA59954.1 nucleolin	736	0.0	G
dbj BAC03738.1 unnamed protein product [Homo sapiens]	712	0.0	UG
ref XP_614626.2 PREDICTED: similar to nucleolin-related prot...	702	0.0	UG
ref NP_072143.1 nucleolin-related protein [Rattus norvegicus...	701	0.0	UG
ref XP_850477.1 PREDICTED: similar to nucleolin-related prot...	681	0.0	G
ref XP_861643.1 PREDICTED: similar to nucleolin-related prot...	678	0.0	G
ref XP_861613.1 PREDICTED: similar to nucleolin-related prot...	678	0.0	G
sp P08199 NUCL_MESAU Nucleolin (Protein C23)	654	0.0	UG
ref NP_036881.1 nucleolin [Rattus norvegicus] >sp P13383 NUC...	643	0.0	UG
gb AAH85751.1 Nucleolin [Rattus norvegicus]	642	0.0	UG
ref XP_861582.1 PREDICTED: similar to nucleolin-related prot...	642	0.0	G
gb AAA36286.1 nucleolin, C23	631	0.0	UG
ref U01018 nucleolin - rat	630	0.0	UG
dbj BAC27474.1 unnamed protein product [Mus musculus]	637	0.0	UG
gb AAH05460.1 Nucleolin [Mus musculus]	632	2e-179	UG
ref NP_035010.3 nucleolin [Mus musculus] >sp P09405 NUCL_MOU...	632	2e-179	G
dbj BAE38940.1 unnamed protein product [Mus musculus]	631	4e-179	UG
dbj BAE36484.1 unnamed protein product [Mus musculus]	631	4e-179	UG
dbj BAE40448.1 unnamed protein product [Mus musculus] >dbj B...	631	5e-179	UG
dbj BAC26311.1 unnamed protein product [Mus musculus]	628	3e-178	UG

The E-Values

- ✓ E-value means *expectation value*
- ✓ The E-value is the measure most commonly used for estimating sequence similarity
- ✓ How many times is a match at least as good expected to happen by chance ?
 - This estimate is based on the similarity measure
- ✓ If a match is highly unexpected, it probably results from something other than chance
 - Common origin is the most likely explanation
 - This is how homology is inferred

Which Value for Your E-Values ?

- ✓ Low E-value \leftrightarrow good hit
 - 1 = bad e-Value
 - 10^{-3} = borderline E-value
 - 10^{-4} = OK E-value
 - 10^{-8} = good E-value
 - 10^{-10} = very good E-value
- ✓ E-values lower than 10^{-8} indicate homology pretty confidently
- ✓ E-values lower than 10^{-6} indicate possible homology
- ✓ E-values higher than 10^{-4} require extra evidence to support homology

Why Use E-Values?

- ✓ E-values make it possible to compare alignment of different lengths
- ✓ E-values are used by most sequence comparison programs
 - PSI-BLAST
 - Domain Search
 - FASTA
- ✓ E-values always have the same meaning
 - You can compare the output of different programs
 - However, be careful - strictly speaking an E-value depends on the database that is searched (its composition - because this influences the chance of finding similar-looking sequences to your query at random, i.e. in a non-homologous sequence)

The Alignments

- ✓ Look for clusters of identity
- ✓ Gray residues are low-complexity regions
- ✓ Grayed-out regions have been removed from your sequence to avoid false hits

Score = 784 bits (2025), Expect = 0.0
 Identities = 707/707 (100%), Positives = 707/707 (100%), Gaps = 0/707 (0%)

```

Query 1  MWLAKAGHTHGEAKKMAAPPPKveedsedeemsedeedsageeevIFQkqkkatrp 60
          MWLAKAGHTHGEAKKMAAPPPKVEEDSEDEEMSEDEDDSSGEEVVIQKQKKAITTP
Sbjct 1  MWLAKAGHTHGEAKKMAAPPPKVEEDSEDEEMSEDEDDSSGEEVVIQKQKKAITTP 60

Query 61  akivvSQkkaavtpakkaavtpgkka:PAKKNITPAKVIPTPGKGAQAQALVPE 120
          AKIVVVSQTKAAVPTPAKKAAVTPGKKAIVTPAKKNITPAKVIPTPGKGAQAQALVPE
Sbjct 61  AKIVVVSQTKAAVPTPAKKAAVTPGKKAIVTPAKKNITPAKVIPTPGKGAQAQALVPE 120

Query 121  tpgkkaatpangakngknakdeddeddeddeddeddeFEFPIVGVKPA 180
          TPGKKAATPANGAKNGKNAKEDSDEDEDEDEDDSDDEDEDEFEFPIVGVKPA
Sbjct 121  TPGKKAATPANGAKNGKNAKEDSDEDEDEDEDDSDDEDEDEFEFPIVGVKPA 180

Query 181  kaapaasadeddeddeddeddeddeeeVMEITTAGKKTIPAKVPMGAKSVA 240
          KAAPAASASEDEDEDEDEDEDEDEDEEEVMEITTAGKKTIPAKVPMGAKSVA
Sbjct 181  KAAPAASASEDEDEDEDEDEDEDEDEEEVMEITTAGKKTIPAKVPMGAKSVA 240

Query 241  eeedeeddeddeddeddeddeeeVVAAGGKXKEMTKQKspeakkqKV 300
          EEEDDEEDEDDEDEDEDEDEDEDEEEVVAAGGKXKEMTKQKspeakkqKV
Sbjct 241  EEEDDEEDEDDEDEDEDEDEDEDEEEVVAAGGKXKEMTKQKspeakkqKV 300
    
```

BLASTing DNA Sequences

- ✓ The BLAST program you need depends on your DNA sequence
 - Coding DNA
 - Non Coding DNA

- ✓ BLASTing DNA sequences is less accurate than BLASTing protein sequences

- ✓ If your sequence is coding, blastx and tblastx will translate it for you on its 6 possible reading frames

BLASTing DNA Sequences

Program	Query	Database
blastn	nucleotide	nucleotide
blastx	nucleotide 	protein
tblastx	nucleotide 	nucleotide

Asking the Right Question with BLAST

Choosing the right flavor of BLAST for DNA

<i>Question</i>	<i>Answer</i>
Am I interested in non-coding DNA?	Yes: use blastn . Never forget that blastn is only for closely related DNA sequences (more than 70 percent identical)
Do I want to discover new Proteins?	Yes: use tblastx .
Do I want to discover proteins encoded in my query DNA sequence?	Yes: use blastx
Am I unsure of the quality of my DNA?	Yes: use blastx if you suspect your DNA sequence is coding for a protein but that it may contain sequencing errors.

Some Reasons for Changing the Default Parameters

<i>Reason</i>	<i>Parameters to Change</i>
The sequence you're interested in contains many identical residues; it has a biased composition.	Sequence filter (automatic masking)
BLAST doesn't report any results.	Change the substitution matrix or the gap penalties.
Your match has a borderline E-value.	Change the substitution matrix or the gap penalties to check the match robustness.
BLAST reports too many matches.	Change the database you're searching OR filter the reported entries by keyword OR increase the number of reported matches OR increase Expect, the E-value threshold OR reject sequences too similar to the query (very low E-values).

PSI-BLAST

- ✓ PSI-BLAST is *Position-Specific Iterated* BLAST
 - More sensitive than BLAST: finds matches BLAST would not find
 - More specific than BLAST: reports fewer false matches
 - A bit slower than BLAST

- ✓ PSI-BLAST finds remote homologues
 - Will let you identify very distant members of your protein family

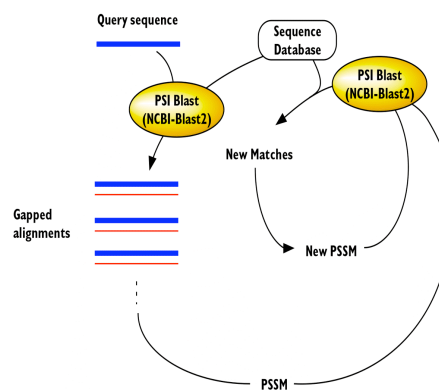
- ✓ PSI-BLAST uses the results of each iteration to increase its specificity

PSI-BLAST Iterations

- ✓ PSI-BLAST uses the best results of the first iteration to build a profile (PSSM)

- ✓ PSI-BLAST uses the profile to re-scan the database

- ✓ PSI-BLAST keeps re-scanning until it stops finding new matches



Some Tips for Using PSI-BLAST

- ✓ If your protein is multi-domain, search one domain at a time
- ✓ PSI-BLAST is slower than normal BLAST because of the iterations
- ✓ You can feed PSI-BLAST with your own PSSM
 - Use the NCBI server for this purpose

Going Farther

- ✓ Each BLAST online server is unique
- ✓ Shop around to find the right database
- ✓ If you need to look for exact matches between a sequence and a genome use BLAT
 - No it's not a typo
 - You can find it at genome.ucsc.edu
- ✓ If you want something more accurate than BLAST, use Smith and Waterman
 - It's also slower than BLAST
 - You can find it at www-btls.jst.go.jp