"Google is arguably one of the greatest inventions. The search engine is one of the greatest inventions in human history."
-- Franklin Foer

"Probabilistic models of transcriptomic (dys)regulation"

David Knowles

ECCR265

Thursday, February 22, 2018

Concepts of Sequence Similarity Searching

- The premise:

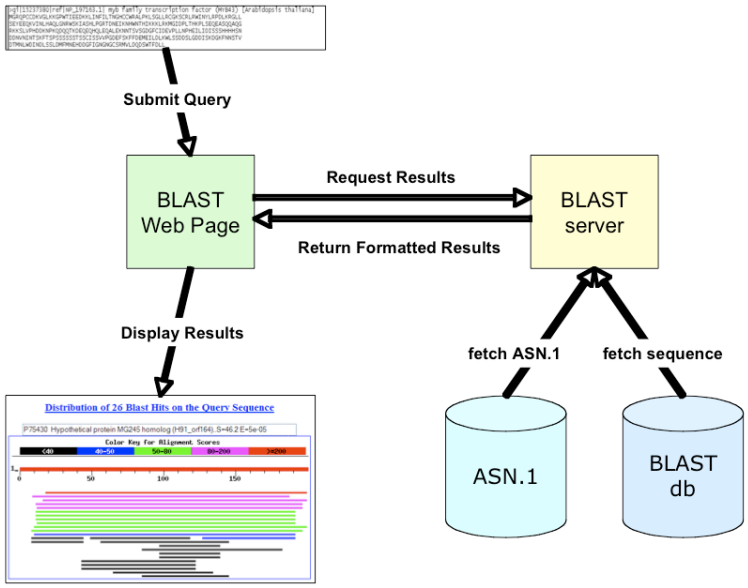
One sequence by itself is not informative; it must be analyzed by comparative methods against existing sequence databases to develop hypothesis concerning relatives and function.

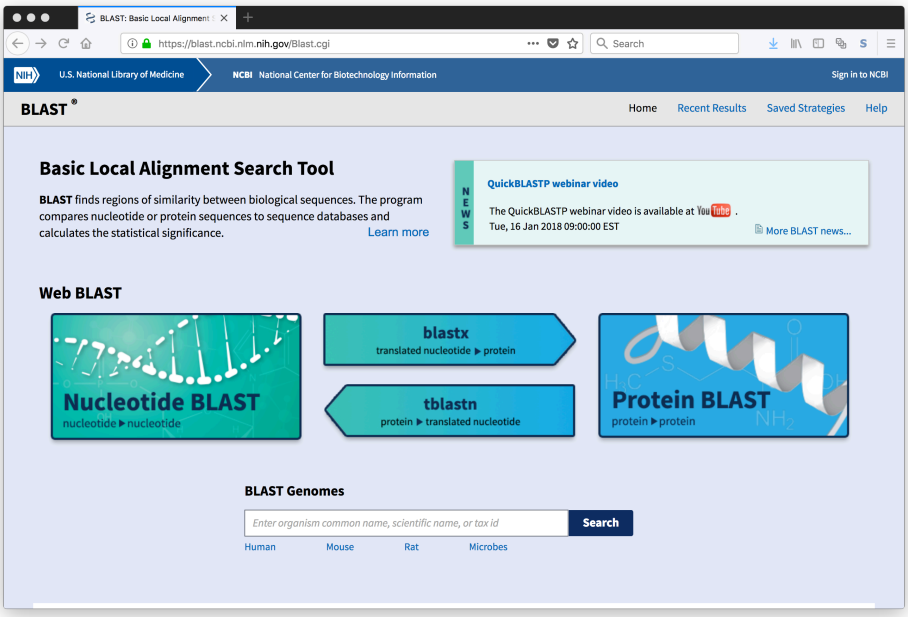
3

The BLAST algorithm

The BLAST programs (Basic Local Alignment Search Tools) are a set of sequence comparison algorithms introduced in 1990 that are used to search sequence databases for optimal local alignments to a query.

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.
- Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." NAR 25:3389-3402.


Submit Query
BLAST Web Page
Request Results
Return Formatted Results
BLAST server
Display Results
fetch ASN.1
fetch sequence
ASN.1
BLAST db
Distribution of 26 Blast Hits on the Query Sequence


BLAST: Basic Local Alignment
https://blast.ncbi.nlm.nih.gov/Blast.cgi
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
BLAST
Home
Recent Results
Saved Strategies
Help
Basic Local Alignment Search Tool
BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.
Learn more
QuickBLASTP webinar video
The QuickBLASTP webinar video is available at YouTube.
Tue, 16 Jan 2018 09:00:00 EST
More BLAST news...
Web BLAST
Nucleotide BLAST
nucleotide ▶ nucleotide
blastx
translated nucleotide ▶ protein
tblastn
protein ▶ translated nucleotide
Protein BLAST
protein ▶ protein
BLAST Genomes
Enter organism common name, scientific name, or tax id
Search
Human
Mouse
Rat
Microbes

BLAST programs

Program	Description
blastp	Compares an amino acid query sequence against a protein sequence database.
blastn	Compares a nucleotide query sequence against a nucleotide sequence database.
blastx	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.
tblastn	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
tblastx	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

BLAST is an experiment!

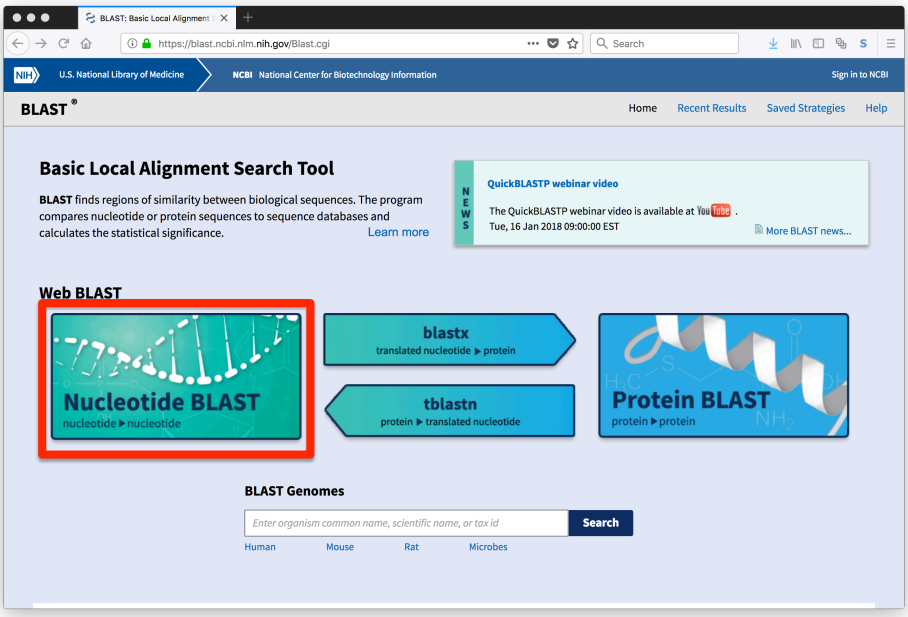
- BLAST is not Google
- BLAST is like doing an experiment: to get good, meaningful results, you need to optimize the experimental conditions
    1. What kind of BLAST?
    2. Pick an appropriate database
    3. Pick the right algorithm
    4. Choose parameters

Step 1:
Nucleotide BLAST vs. protein BLAST

- Largely determined by your query sequence

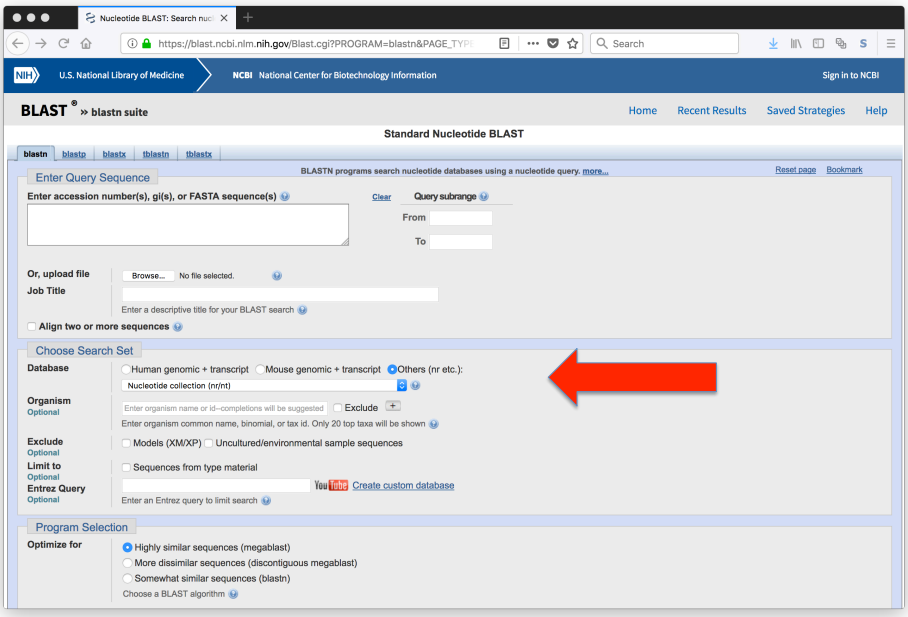
BUT

- If your nucleotide sequence can be translated to a peptide sequence, you probably want to do it
- Protein blasts are more sensitive and biologically significant
- Sometimes it makes sense to use other blasts

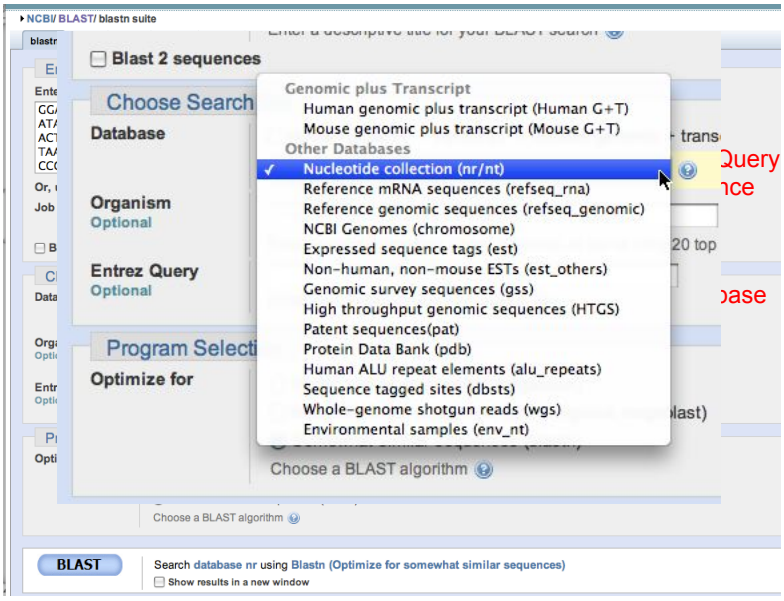


Step 2: Choose a Database

- Too large:
  - Takes longer
  - Too many results
  - More random, meaningless matches
- Too small or wrong one:
  - Miss significant matches

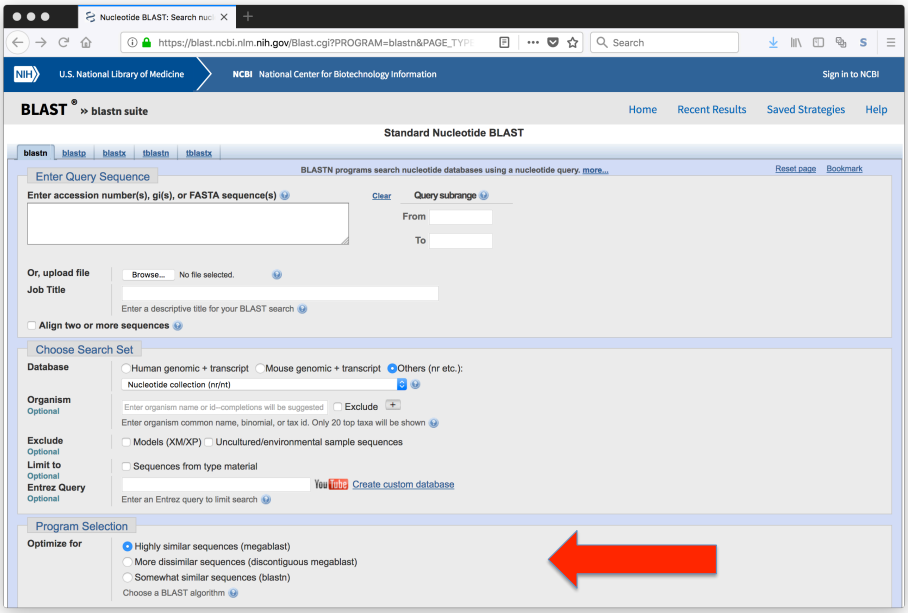


blastn Options



Protein Databases

- Non-redundant protein sequences (nr)
  - Kitchen-sink:
    - Translations of GenBank coding sequences (CDS)
    - RefSeq Proteins
    - PDB (RCSB Protein Data Bank - 3d-structure)
    - SwissProt
    - Protein Information Resource (PIR)
    - Protein Research Foundation (Japanese DB)
- Reference proteins (refseq_protein)
  - NCBI Reference Sequences: Comprehensive, integrated, non-redundant, well-annotated set of sequences
- Swissprot protein sequences (swissprot)
  - Swiss-Prot: European protein database (no incremental updates)



Step 3:
Choose an Algorithm

- How close a match are you looking for?
- Determines how similarities are “scored”
- Affects speed of search and chance of missing match
- Again, what is the goal of the search?

more BLAST programs

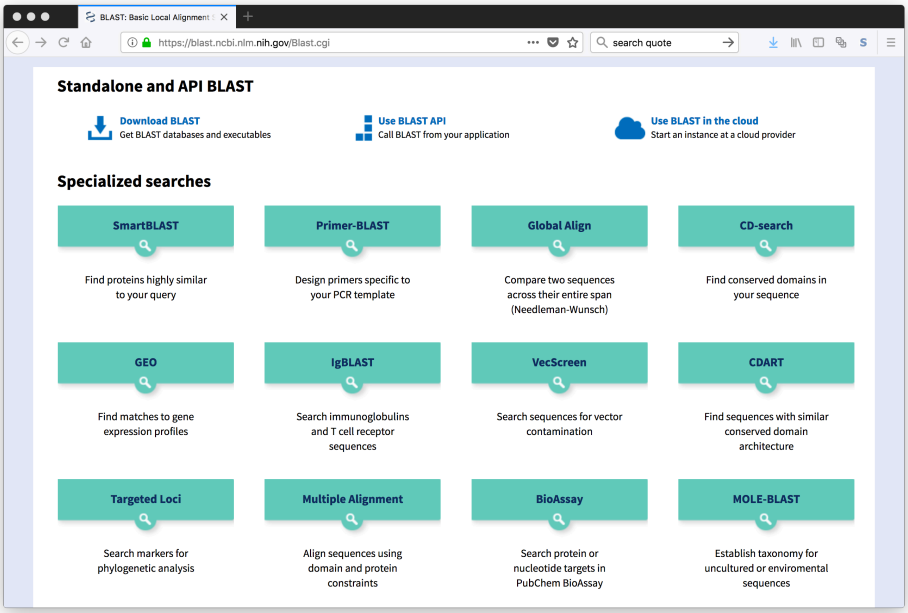
Program		Notes
Megablast	Contiguous	Nearly identical sequences
	Discontiguous	Cross-species comparison
Position Specific	PSI-BLAST	Automatically generates a position specific score matrix (PSSM)
	RPS-BLAST	Searches a database of PSI-BLAST PSSMs

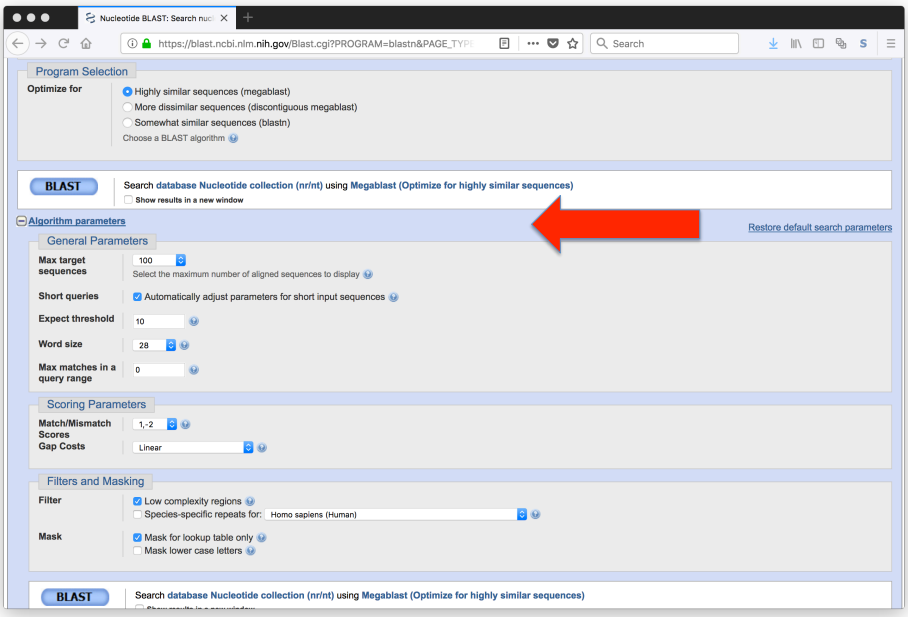
nucleotide only

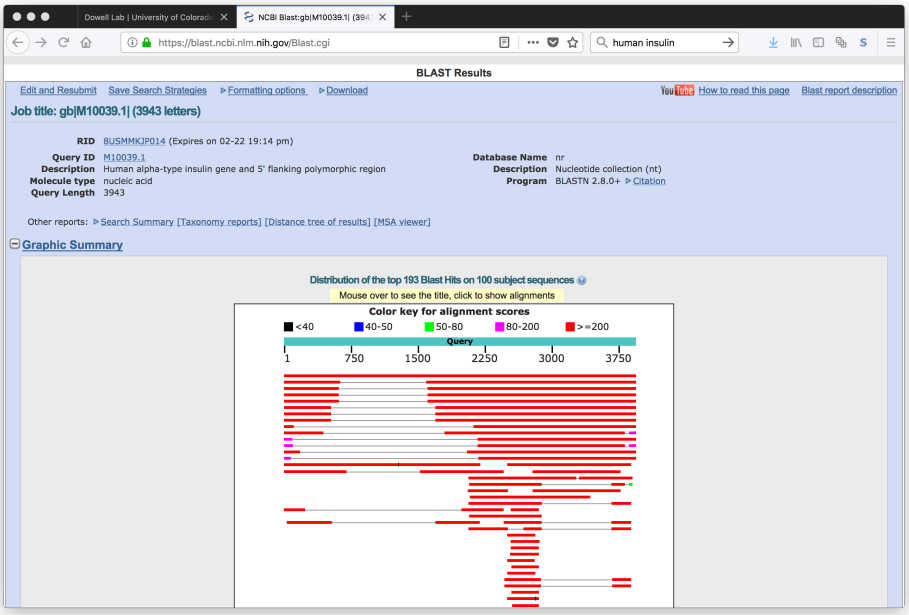
protein only

Specialized BLAST versions

- **Megablast** – (nucleotides) very fast for highly similar sequences
- **PSI-Blast** – (proteins) Position Specific Iterated (profile searching)
- Whole slew of even more specialized BLAST versions for specific reoccurring cases …


BLAST: Basic Local Alignment
https://blast.ncbi.nlm.nih.gov/Blast.cgi
search quote
Standalone and API BLAST
Download BLAST
Get BLAST databases and executables
Use BLAST API
Call BLAST from your application
Use BLAST in the cloud
Start an instance at a cloud provider
Specialized searches
SmartBLAST
Find proteins highly similar to your query
Primer-BLAST
Design primers specific to your PCR template
Global Align
Compare two sequences across their entire span (Needleman-Wunsch)
CD-search
Find conserved domains in your sequence
GEO
Find matches to gene expression profiles
IgBLAST
Search immunoglobulins and T cell receptor sequences
VecScreen
Search sequences for vector contamination
CDART
Find sequences with similar conserved domain architecture
Targeted Loci
Search markers for phylogenetic analysis
Multiple Alignment
Align sequences using domain and protein constraints
BioAssay
Search protein or nucleotide targets in PubChem BioAssay
MOLE-BLAST
Establish taxonomy for uncultured or enviromental sequences

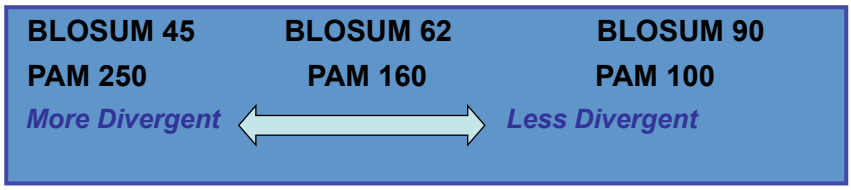

Nucleotide BLAST: Search nuc
https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYP
Search
Program Selection
Optimize for
Highly similar sequences (megablast)
More dissimilar sequences (discontiguous megablast)
Somewhat similar sequences (blastn)
Choose a BLAST algorithm
BLAST
Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
Show results in a new window
Algorithm parameters
Restore default search parameters
General Parameters
Max target sequences
100
Select the maximum number of aligned sequences to display
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
10
Word size
28
Max matches in a query range
0
Scoring Parameters
Match/Mismatch Scores
1,-2
Gap Costs
Linear
Filters and Masking
Filter
Low complexity regions
Species-specific repeats for:
Homo sapiens (Human)
Mask
Mask for lookup table only
Mask lower case letters
BLAST
Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)



Where does the score (S) come from?

- The quality of each pair-wise alignment is represented as a score and the scores are ranked.
- Scoring matrices are used to calculate the score of the alignment base by base (DNA) or amino acid by amino acid (protein).
- The alignment score will be the sum of the scores for each position.

BLOSUM vs PAM



- BLOSUM 62 is the default matrix in BLASTp 2.0. Though it is tailored for comparisons of moderately distant proteins, it performs well in detecting closer relationships. A search for distant relatives may be more sensitive with a different matrix.

Database searching: What is a good hit?

- Say that you search the non-redundant protein database at NCBI, containing roughly one million sequences. What p-value threshold should you use?

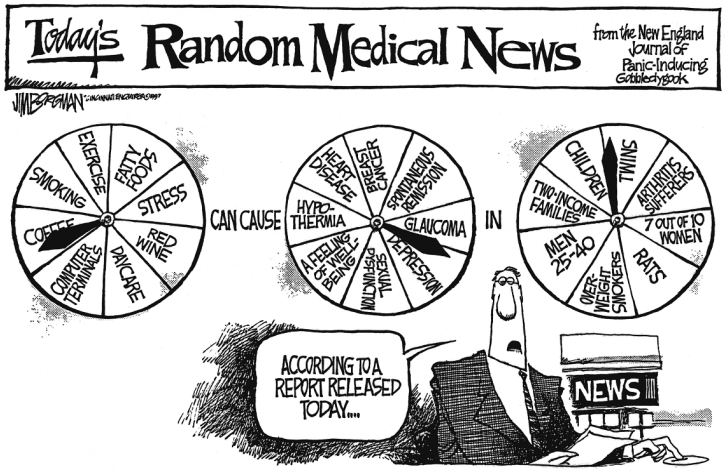
Multiple testing

- Say that you perform a statistical test with a 0.05 p-value threshold, but you repeat the test on twenty different observations.
- Assume that all of the observations are explainable by the null hypothesis.
- What is the chance that at least one of the observations will receive a p-value less than 0.05?

Multiple testing

- Say that you perform a statistical test with a 0.05 threshold, but you repeat the test on twenty different observations. Assuming that all of the observations are explainable by the null hypothesis, what is the chance that at least one of the observations will receive a p-value less than 0.05?
- Pr(making a mistake) = 0.05
- Pr(not making a mistake) = 0.95
- Pr(not making any mistake) = 0.95^{20} = 0.358
- Pr(making at least one mistake) = 1 - 0.358 = 0.642
- There is a 64.2% chance of making at least one mistake.

Failure to correct for multiple hypothesis testing:



Bonferroni correction

- Assume that individual tests are *independent*.
  (Is this a reasonable assumption?)
- Divide the desired p-value threshold by the number of tests performed.
- For the previous example, 0.05 / 20 = 0.0025.

  Pr(making a mistake) = 0.0025

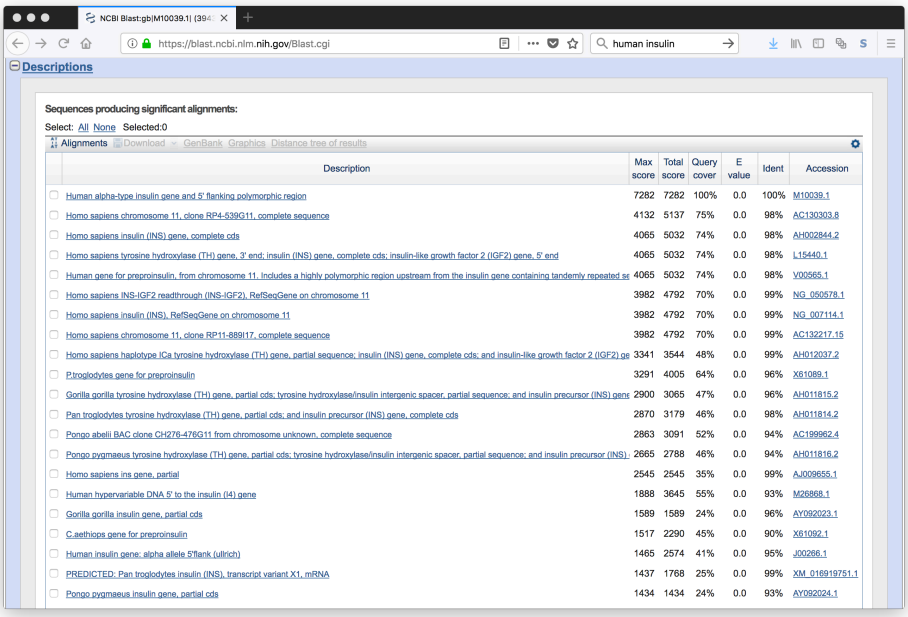
  Pr(not making a mistake) = 0.9975

  Pr(not making any mistake) = 0.9975^{20} = 0.9512

  Pr(making at least one mistake) = 1 - 0.9512 = 0.0488

Database searching

- Say that you search the non-redundant protein database at NCBI, containing roughly one million sequences. What p-value threshold should you use?
- Say that you want to use a conservative p-value of 0.001.
- Recall that you would observe such a p-value by chance approximately every 1000 times in a random database.
- A Bonferroni correction would suggest using a p-value threshold of 0.001 / 1,000,000 = 0.000000001 = 10^{-9}.



Sequence Similarity Searching – The statistics are important

- Discriminating between real and artifactual matches is done using an estimate of probability that the match might occur by chance.
- The Score (S) reflects the desired scoring scheme. Karlin-Altshul statistics gives us a sense for whether that's a good score (p-value), but BLAST reports E-values!

E-values

- A p-value is the probability of making a mistake … but properly adjusting it requires *knowing the size of the database!*
- The E-value is the expected number of times that the given score would appear in a random database of the given size.
- One simple way to compute the E-value is to multiply the p-value times the size of the database.

BLAST actually calculates E-values in a much more complex way.

E-values

- Thus, for a p-value of 0.001 and a database of 1,000,000 sequences, the corresponding E-value is 0.001 × 1,000,000 = 1,000.
- When E < 0.01, e-value and p-value are essentially the same.
- Hence smaller numbers are more significant!

Short sequences can't get good e-values!!

- What is the probability of finding a 12 base fragment in a "random" genome?

  4^{12} = 16,777,216 (once per 16 million bases)

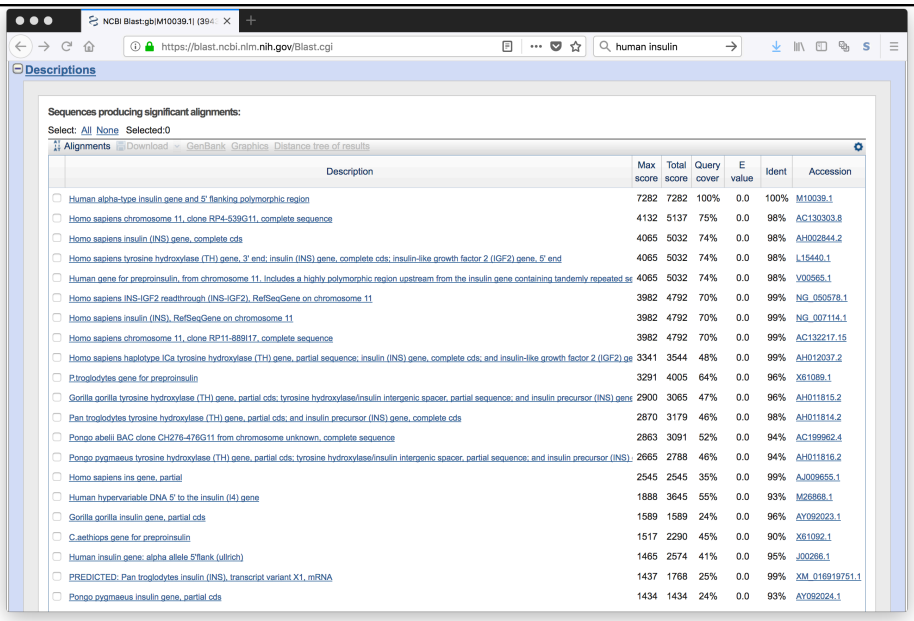
- What length DNA fragment is needed to define a unique location in the human genome?

  4^{16} = 4,294,967,296 (4 billion bases)

What do the Score and the e-value really mean?

- The quality of the alignment is represented by the Score (S).
  - The score of an alignment is calculated as the sum of substitution and gap scores. Substitution scores are given by a look-up table (PAM, BLOSUM) whereas gap scores are assigned empirically .
- The significance of each alignment is computed as an E value (E).
  - Expectation value. The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.

The E-value is not a probability; it's an expected value!



It is important to keep in mind that the E() value does not represent a measure of similarity between the two sequences.

E-values

- Statistical significance depends on both the size of the alignments and the size of the sequence database
  - Important consideration for comparing results across different searches
  - E-value increases as database gets bigger
  - E-value decreases as alignments get longer

Interpretation of E-values

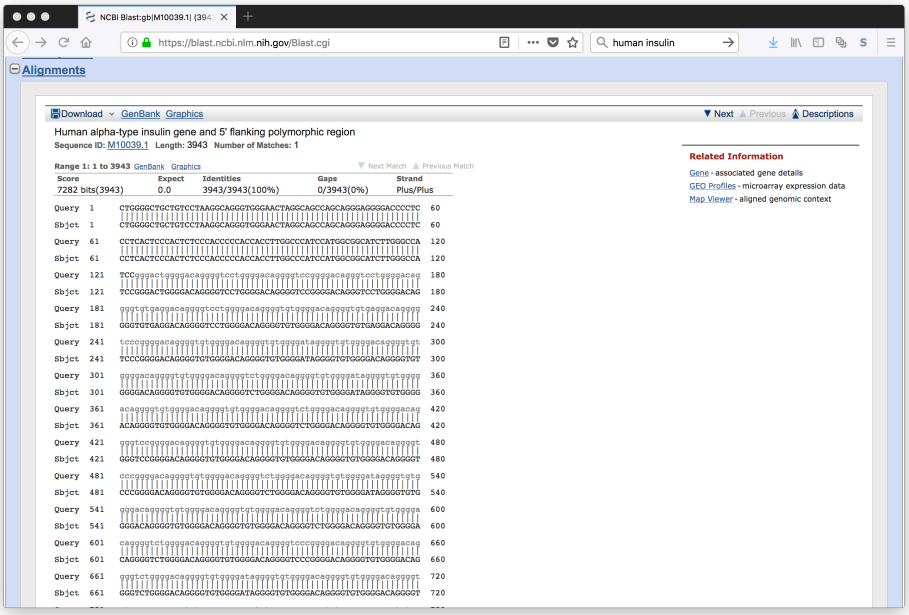
- very low E() values are identical genes or possibly closely related genes
- moderate E() values are almost certainly related genes
- long list of gradually declining of E() values indicates a large gene family
- long regions of moderate similarity are more significant than short regions of high identity

Homology: Some Guidelines

- Similarity can be indicative of homology
- Generally, if two sequences are significantly similar over entire length they are likely homologous
- BUT ⋯ Low complexity regions can be highly similar without being homologous
- AND ⋯ Homologous sequences not always highly similar

Biological Relevance

- It is up to you, the biologist to scrutinize these alignments and determine if they are significant.
- Were you looking for a short region of nearly identical sequence or a larger region of general similarity?
- Are the mismatches conservative ones?
- Are the matching regions important structural components of the genes or just introns and flanking regions?



>gb|AAL08419.1| PTEN [Takifugu rubripes]
Length=412

Score = 197 bits (501), Expect = 2e-49, Method: Composition-based stats.
Identities = 95/100 (95%), Positives = 98/100 (98%), Gaps = 0/100 (0%)

```
Query  2    IVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI  61
            +VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI
Sbjct  8    MVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI  67

Query  62   YNLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFKQN  101
            YNLCAERHYD AKFNCRVAQYPFEDHNPPQLELIKPF ++
Sbjct  68   YNLCAERHYDAAKFNCRVAQYPFEDHNPPQLELIKPFCED  107
```

Score = 83.6 bits (205), Expect = 4e-15, Method: Composition-based stats.
Identities = 60/103 (58%), Positives = 68/103 (66%), Gaps = 32/103 (31%)

```
Query  99   KQNKMLKKDKMPHFWVNTFFIPGPEEV---------------------------------D  126
            KQNKM+KKDKMPHFWVNTFFIPGPEE                                  +
Sbjct  260  KQNKMMKKDKMPHFWVNTFFIPGPEESRDKLENGAVNNADSQQGVPAPGQGQPQSAECRE  319

Query  127  NDKEYLVLTLTkndldkankdkanRYFSPNFKVKLYFTKTVEE  169
            +D++YL+LTL+KND DKANKDKANRYFSPNFKVKL F+KTVEE
Sbjct  320  SDRDYLILTLSKNDRDKANKDKANRYFSPNFKVKLCFSKTVEE  362
```

>gb|AAH93110.1| Ptenb protein [Danio rerio]
Length=289

Score = 197 bits (500), Expect = 2e-49, Method: Composition-based stats.
Identities = 95/99 (95%), Positives = 98/99 (98%), Gaps = 0/99 (0%)

```
Query  3    VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKIY  62
            VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHK+HYKIY
Sbjct  9    VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKDHYKIY  68

Query  63   NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFKQN  101
            NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPF ++
Sbjct  69   NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFCED  107
```

Taxonomy Reports