BIOINFORMATICS

Lecture 10

Sequence Alignment - III

ROBI POLIKAR

Rowan University

SIGNAL PROCESSING & PATTERN RECOGNITION LABORATORY @ ROWAN UNIVERSITY

http://www.photoshopmonster.com/Make-a-sand-blast-text-effect-image_93.html

Journal of Molecular Biology
Volume 215, Issue 3, 5 October 1990, Pages 403-410

doi:10.1016/S0022-2836(05)80360-2 | How to Cite or Link Using DOI
Permissions & Reprints

Basic local alignment search tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2], Eugene W. Myers[3], David J. Lipman[1]

1 National Center for Biotechnology Information National Library of Medicine, National Institutes of Health Bethesda, MD 20894, U.S.A.

2 Department of Computer Science The Pennsylvania State University, University Park, PA 16802, U.S.A.

3 Department of Computer Science University of Arizona, Tucson, AZ 85721, U.S.A.

Received 26 February 1990; Accepted 15 May 1990. Edited by S. Brenner. Available online 6 February 2007.

A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straight-forward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than existing sequence comparison tools of comparable sensitivity.

BLAST

- Basic Local Alignment Search Tool is perhaps the most widely used bioinformatics algorithms, not only because it solves the most commonly need operation – sequence alignment – but also because it does so very rapidly
- BLAST emphasizes speed over sensitivity, however, the sensitivity lost is relatively a small price paid for the incredible speed gained over "optimal" algorithms.
- BLAST is also faster than FASTA with comparative sensitivity, and without many of the shortcomings mentioned earlier.
- Note however that BLAST is a heuristic algorithm, and just like FASTA, is not guaranteed to find the optimal alignment.

BLAST

- BLAST, just like FASTA, also starts by finding short words, referred to as k-tuples in FASTA and k-mers in BLAST, that are common between the query and database sequences.
- However, BLAST differs from FASTA in two important aspects:
  1. While FASTA looks for k-tuples that are identical (exact match) between the query and target sequence, BLAST looks for k-mers in the target that score above a certain threshold T, when aligned with the query sequence. Hence, k-mers need not be identical in BLAST, which solves many of the important shortcomings of FASTA.
  2. While FASTA uses hashing and chaining to find / locate the k-tuples that match, BLAST uses a finite state machine (FSM), which identifies and locates certain subsequences in a long string of letters.
- The initial high-scoring ungapped alignments found by BLAST are called ***high scoring segment pairs (HSPs)***, the highest scoring of which is called the ***maximal segment pair (MSP)***.
  - BLAST's MSP corresponds to FASTA's *init1* diagonal

BLAST
PARAMETERS

- Determine the user defined parameters:
  - **Word size, *k*** in k-mer: What length of short sequences should BLAST search for?
    - Typically 3 (2~4) for AA sequences, 11 for DNA sequences.
  - **Threshold *T*** : Only those k-mer matches (possibly not exact) that score above T are considered for "seeding" the algorithm.
    - Typically an integer between 11 and 19.
  - **Drop-off X**: The amount of drop in score that will stop the sequence extension
    - Typically around 20.
  - **Scoring / substitution matrix**
    - Typically BLOSUM62. Choose based on the expected evolutionary distance.
  - Others
    - **Gap penalties**

SIDEBAR

FINITE STATE MACHINES

- A finite state machine, also called a finite state automata, is a computational algorithm that can be used to model a variety of phenomena that are described by a sequence of states that control the behavior of a system.
- Each state is a particular behavior or output, the system provides in response to a particular input.
  - In a FSM, given the current state and an input, the next state of the FSM can be deterministically determined, unlike a hidden Markov Model, where the current state and the input gives only a probabilistic information on what the next state may be.
  - Given a new input, the response of the FSM may be emission of a particular response, as well as transitioning to a new state.
- In the context of bioinformatics, FSM is used to identify a (relatively short) pattern of strings in a long sequence.
  - The inputs are each subsequent residue, whereas the states are typically subsequences of the pattern being sought

FINITE STATE MACHINES

- Given a list of k-mers from the target sequence as an input, the FSM can produce all possible matches (possibly not exact, but that exceed a threshold T) in a very efficient and fast way.
- For example, consider a 3-mer query CHH. If T is set to 19, based on the BLOSUM 62 matrix, CHY and CYH – in addition to the perfect match CHH – will generate scores that exceed 19.
  - So want to build an FSM that will identify and alert anytime any of the 3-mers CHH, CHY or CYH appears in the sequence
  - Side note: T=14 was a common threshold used in early versions of BLAST. Note that with this threshold, 3-mers that consists of A, I, L or V can never reach the threshold.

CHH – CHH: 9+8+8 = 25
CHH – CHY: 9+8+2 = 19
CHH – CYH:9+2+8 = 19

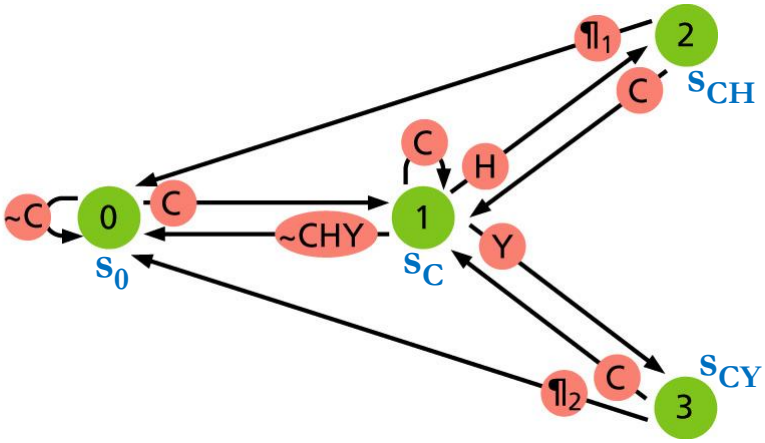
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5							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

FINITE STATE MACHINES

EXAMPLE

- Let's use the same example: We have a 3-mer query CHH. T is set to 19, and based on the BLOSUM 62 matrix, CHH, CHY and CYH pass the threshold. We need a mechanism to scan the database sequence and output the above 3-mers when they are encountered.
  - We have the transition diagram and a table to track the outputs. We start at state 0 (s_0).
  - Anything other than a C (indicated as ~C) returns us to state 0, since all k-mers start with C.
  - A "C" takes us to state 1, which we can also call state s_C. Then an "H" takes us to state 2 (s_{CH}), the only way this state can be reached, requiring an input sequence of CH
  - A "Y" input seen at state 1 takes us to state 3 (s_{CY}), the only way this state can be reached, requiring an input sequence of CY
  - A second "C" at state s_C, returns us to the same state, as that can now be considered as the beginning of the new sequence starting with C.

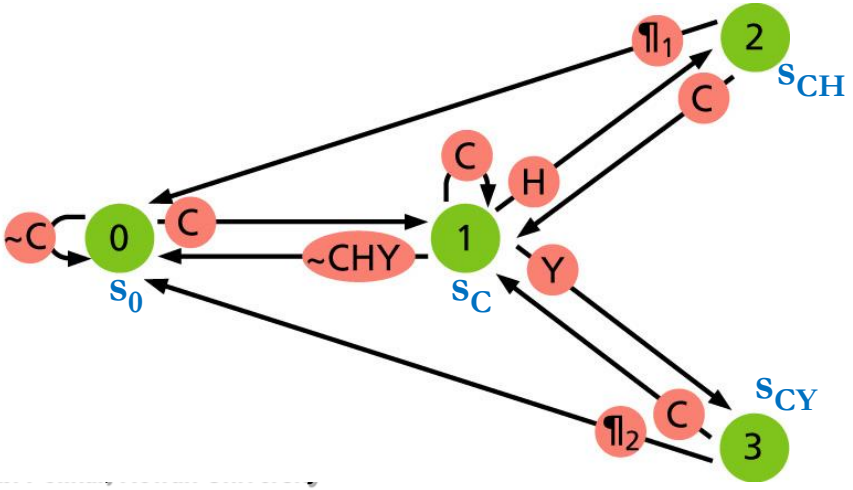
	INPUT	OUTPUT
$\P_1$:	~CHY	none
	H	CHH
	Y	CHY
$\P_2$:	~CH	none
	H	CYH



FINITE STATE MACHINES
EXAMPLE

- Continuing with tracing the FSM for this example:
  - From states 2 and 3 (s_{CH} and s_{CY}), we go back to state 0 (start new search), unless the next input is C:
    - **If it is C, we go back to state 1, s_C, as that means we start a new search that starts with C**
  - From s_{CH} : If input is "H" → Return to state 0 and output the identified 3-mer CHH
    If input is "Y" → Return to state 0 and output the identified 3-mer CHY
    If input is "C" → Return to state 1 s_C for a new sequence that starts with C
  - From s_{CY}: If input is "H" → Return to state 0 and output the identified 3-mer CYH
    If input is "Y" → Return to state 0. No output is provided.
    If input is "C" → Return to state 1 s_C for a new sequence that starts with C

	INPUT	OUTPUT
$\P_1$:	~CHY	none
	H	CHH
	Y	CHY
$\P_2$:	~CH	none
	H	CYH

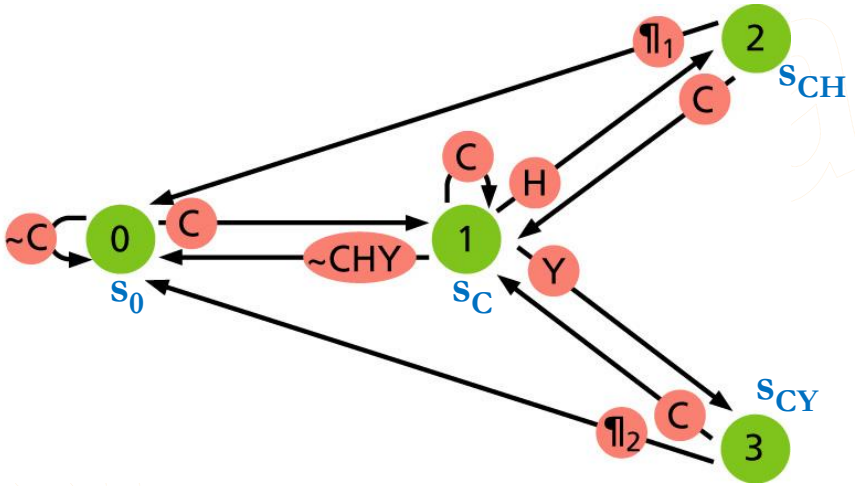


FINITE STATE MACHINES

EXAMPLE

- If the input is CHCYHC → the states visited are 0121301, where the return to state 0 is also accompanied by the sequence CYH identified as a "hit"

	INPUT	OUTPUT
¶1:	~CHY	none
	H	CHH
	Y	CHY
¶2:	~CH	none
	H	CYH



- Note that in reality, with 20^3 = 8000 3-mers, the actual FSM diagram can be quite complex. To minimize the size, we make sure that whenever a residue forces the algorithm to start a new search, that search uses as much of the identified sequence as possible by returning not to the initial state 0, but to an intermediate state that can use that partial sequence.
- Note that the FSMs are created only for the k-mers that are found in the query sequence.

BLAST

OVERALL APPROACH

- Given a query sequence and a database of potential targets:
- For each target
  - Step 1: Obtain the list of words in the target sequence (k-mers), that give a score of *T* or higher when aligned with the query sequence. This is called ***seeding***.
  - Step 2: Scan the database for ***hits*** with the compiled list of words obtained in Step 1
    - **Finite State Machines designed to catch the determined hits in the database sequence.**
  - Step 3: Extend the hits to form high scoring segment pairs. The extension proceeds in both directions from the k-mer until the score drops by more than X compared to current best score.
  - Step 4: Find the highest scoring segment (the maximal segment pair) or those whose score exceeds (another user set) threshold *S*.
  - Step 5: If needed – combine two or more HSP regions into a longer alignment.
  - Step 6: Evaluate the statistical significance of the alignments / scores that exceed the threshold.
  - Report every match – across the database – whose "Expect Score, E" is lower than a threshold of significance.

BLAST
STEP 1: SEEDING

- Obtain the list of words in the target sequence (k-mers), that give a score of *T* or higher when aligned with the query sequence.
  - Assume that the query sequence is **P Q G E F G**
  - We have the following four 3-mers (recall, the number of k-mers is always N-k+1:

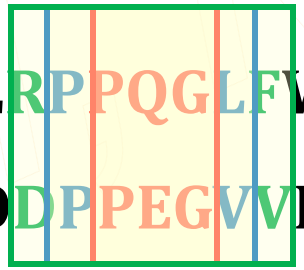
P Q G
Q G E
G E F
E F G

Each of these 3-mers are then scored against each and every one of the k-mers in each of the target sequence. For long sequences, this could well include all 8000 posisble k-mers

  - So, of all pairwise scorings for P Q G (using the BLOSUM-62 matrix), we may find the following high scoring ones:
    - **P Q G (of course, this is a perfect match)→ score of 7+5+6 = 18**
    - **P E G → score of 7+2+6 = 15**
    - **P Q A→ score of 7+5+0 = 12**
  - If T is set to 13, then PQG and PEG will be chosen as hits, where PQA will be dropped

STEP 2&3: SCANNING & EXTENSION

- Scan the database for hits with the compiled list of words obtained in Step 1
  - We are now looking for the locations of these matches in the database
  - Finite State Machines designed to catch the determined hits in the database sequence.
- Once the hits are located both in the query and the target sequence, extend the hits to form ***high scoring segment pairs***.
  - The extension proceeds in both directions from the k-mer until the score drops by X or more compared to current best score. Let's take X = 3 (a typical value is 20)



7+2+6 = 15

7+7+2+6+1 = 23 → High scoring segment - HSP

-2+7+7+2+6+1-1 = 20

BLAST
STEP 4&5: FIND MSP OR HSPS

- Steps 4 & 5: Find the highest scoring segment (the ***maximal segment pair***) or those whose score exceeds (another user set) threshold *S*.
  - The MSP or all those HSPs whose score exceed S are then considered for final listing.
  - S can be determined again using random sampling: Examine the distribution of all alignment scores against those of random sequences, and determine a large enough S to guarantee that all HSPs scoring higher than S are significant.
  - Typically, S is set such that approximately only 2% of the database sequences will have an HSP of greater score
  - No attempt is made by (this original version of) BLAST.

GAPPED BLAST
BLAST 2

- The newer version of BLAST, also called Gapped Blast or Blast 2, uses a so-called two-hit method to save more time.
  - The underlying assumption is that any biologically significant alignment is likely to have at least two high-scoring hits that fall on the same diagonal of the scoring matrix.
  - So Step 1 of the algorithm is slightly adjusted to search for the hits that fall on the same diagonal within a given distance, typically 40 residues.
    - This normally reduced the total number of possible matches, as it is more difficult to find two hits that fall on a diagonal, then finding only one.
    - To maintain a similar sensitivity, and not to miss any hits, the threshold T is reduced, typically to 11 (from 14 or higher), which generates more initial hits.
  - Only a few of these hits will be close enough to each other on the same diagonal.
    - The second hit of such pairs is extended, first ungapped. Then, the HSPs that exceed a threshold S are determined as in regular BLAST.
  - Alignments whose scores exceed S are then used to seed a dynamic programming based optimal local alignment tool to find the gapped alignment.
    - BLAST2 is approximately 3 times faster than ungapped BLAST.

STEP 6: EVALUATING SIGNIFICANCE

- Finally, we evaluate the statistical significance of the alignments / scores that exceed the threshold.
  - Q: Is the alignment score obtained significantly higher than one would expect for two unrelated sequences?
  - A: We need to know the specific distribution that alignment scores follow to be able to compute this.
    - If alignment scores were normally distributed, we could compute the mean / sdt.dev. And use the Gaussian tables to determine the exact probabilities of observing the scores.
    - Alignment scores are NOT normally distributed, because we always get the best possible score for a given pair of sequences, hence the scores obtained are always on the extreme end of the overall distribution of scores.
- Rigorous theoretical analyses have shown that optimal local alignment scores follow ***Gumbel extreme value distribution (EVD)*** for ungapped alignments. For gapped alignments, this distribution is approximate.

$$\mu = \frac{\log(Kmn)}{\lambda}$$

EXTREME VALUE DISTRIBUTION

- Given two sequences of length m and n, the mean of EVD (i.e., where it peaks) is given as

$$\mu = \frac{\log(Kmn)}{\lambda}$$

where λ and K are constants whose values depend on the substitution matrix.

  - λ is a normalization factor of the substitution matrix, and is computed as the solution of

$$\sum_{A,B} p_A p_B e^{s_{A,B}\lambda}$$

where A and B are any two residues, p_A and p_B are the relative frequency of their occurrence in the sequence, and $s_{A,B}$ is the substitution score from the given substitution matrix.

- Then, the probability of the alignment score S being higher than some value x, or alternatively, the probability of finding at least one HSP with a $S \geq x$ is

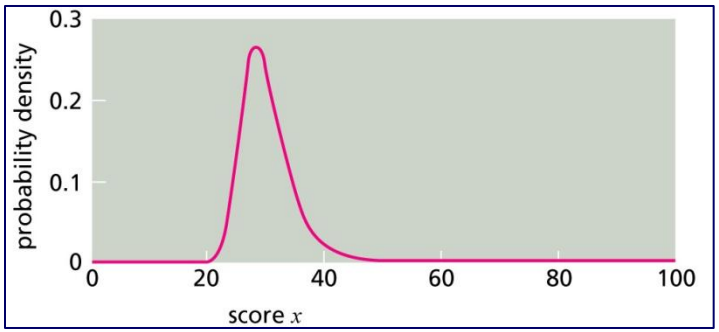
$$P(S \geq x) = 1 - \exp\left(-e^{-\lambda(x-\mu)}\right) = 1 - \exp\left(-e^{-\lambda\left(x - \frac{\log(Kmn)}{\lambda}\right)}\right)$$

$$= 1 - \exp\left(-e^{-(\lambda x - \log(Kmn))}\right) = 1 - \exp\left(-e^{-\lambda x} e^{\log(Kmn)}\right)$$

$$= 1 - \exp\left(-Kmne^{-\lambda x}\right)$$

EXTREME VALUE DISTRIBUTION

- The probability density of the EVD shows a slower (non-symmetric) decay:



$$P(S \geq x) = 1 - \exp\left(-Kmne^{-\lambda x}\right)$$

- If $P(S \geq x) < 0.01$, the alignment is considered significant at the 1% level.
- The above distribution is for $\lambda = 0.286$, K = 0.055, m=n=245 residues. Note that for this set of values, any alignment with a score over 45~50 seems to be significant.
- For gapped alignments, it turns out that the scores still follow EVD, however, there is no theoretical way to determine the λ and K parameters. They are estimated.

A CORRECTION

- In calculating the E-values, BLAST does not use the actual sequence lengths m and n, but rather a modified – specifically reduced – version of these values:
  - BLAST uses ***effective sequence length m' and n'*** to compensate for the edge effect
    - It is unlikely that a true alignment starts near the ends of either the query or target sequence, because there will not be enough sequence to build an optimal alignment.

EXPECTATION VALUE

- Both FASTA and BLAST report an expectation, so called the E-value, which is related to the probability $P(S \geq x)$, but is also related to the number of sequences, D, in the database being searched.
  - Specifically, the E-value is the number of database sequences NOT related to the query sequences that are expected to have an alignment score S greater than the observed score x due to chance.
  - Hence, while the probability is a number between 0 and 1, the E-value is a number between 0 and D. It can be approximated as

$$E \approx 1 - e^{-P(S>x)D} \qquad P(S \geq x) = 1 - \exp\left(-Kmne^{-\lambda x}\right)$$

  - Of course, we want smaller E-values. An E-value of 1 means that we can expect one of the unrelated sequences in the database of D sequences to have an alignment score as high as the one we have obtained with our particular alignment.
    - **E-values that are less than 0.01 are typically considered statistically significant. However, we often look for much smaller E values for long sequences to infer biological significance.**
  - You may also see the E-value defined as the argument of the exponential above: $E = Kmne^{-\lambda S}$. This makes intuitive sense: doubling the length of either sequence doubles the number of HSP attaining the given score S. Note that $P(S>x)$ is the probability of finding at least one HSP with a $S \geq x$. So then, this probability is $P(S>x) = 1\text{-}e^{-E}$, which is the p-value.
  - So p-value (a probability) and E –value (expected # of HSP to have a score greater than S) are not the same things. BLAST returns E-values because, it is easier to interpret. However, for $E<0.01$ $p \sim E$

INTERPRETATION OF E-VALUES

- Very low E-values, e.g., $E < e^{-100}$ typically represent homologs or identical genes
- Moderately low *E*-values, e.g., $E < e^{-50}$ indicate related genes
- If BLAST or FASTA returns many alignments with similar but gradually increasing (or decreasing) E-values, this indicates that the query sequence has matched a large gene family.
- Note that long regions with moderate *E*-values are more significant than short regions with much lower *E*-values or even perfect matches.
  - Q: However, does the statistical significance also indicate biological significance?
  - A: This is open to interpretation and only the analyzing person can make that judgment, based on the length of the matching alignment, whether the mismatches are biologically meaningful (e.g., conserved regions), do the matches actually code genes, etc.
- **Remember:** E-values depend not only on the alignment score, but also on the sequence lengths and the number of sequences in the database being searched.

NUCLEOTIDE ←→PROTEIN ALIGNMENT

- Comparing a nucleotide sequence to a protein sequence is also encountered often:
  - Most protein sequences are nonredundant, and contain only highly reliable sequences that are usually clear of minor variants. Protein sequences are therefore more reliable than nucleotide sequences
    - When we are analyzing a new nucleotide sequence, a better measure of similarity can be obtained by comparing it to a AA sequence than to another NT sequence
  - In order to do this, the nucleotide sequence must first be converted into a protein sequence. There are, however, two problems with this approach:
    - A NT sequence can be translated into protein using one of six open reading frames. Which one shall will align to?
    - Insertion and deletion errors in the NT sequence, which can then cause frameshift in the AA sequence.
  - BLAST (and FASTA) have variations, such as blastx and tblastx that can translate the NT sequence to a protein sequence before the alignment. These algorithms can handle different open reading frames (by analyzing all six), but cannot handle the frameshift.

FAMILY OF BLAST ALGORITHMS

- **Nucleotide-nucleotide BLAST (blastn):** Given a DNA query, returns the most similar DNA sequences from the DNA database that the user specifies.
- **Protein-protein BLAST (blastp)** Given a protein query, returns the most similar protein sequences from the protein database that the user specifies.
- **Position-Specific Iterative BLAST (PSI-BLAST)** This program is used to find distant relatives of a protein. First, a list of all closely related proteins is created. These proteins are combined into a general "profile" sequence, which summarizes significant features present in these sequences. A query against the protein database is then run using this profile, and a larger group of proteins is found. This larger group is used to construct another profile, and the process is repeated. By including related proteins in the search, PSI-BLAST is much more sensitive in picking up distant evolutionary relationships than a standard protein-protein BLAST.
- **Nucleotide 6-frame translation-protein (blastx)** compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.
- **Nucleotide 6-frame translation-nucleotide 6-frame translation (tblastx)** is the slowest of the BLAST family. It translates the query nucleotide sequence in all six possible frames and compares it against the six-frame translations of a nucleotide sequence database. The purpose of tblastx is to find very distant relationships between nucleotide sequences.
- **Protein-nucleotide 6-frame translation (tblastn)** compares a protein query against all six reading frames of a nucleotide sequence database.
- **Large numbers of query sequences (megablast)** concatenates many input sequences together to form a large sequence before searching the BLAST database, then post-analyze the search results to glean individual alignments and statistical values. When comparing large numbers of input sequences via the command-line BLAST, "megablast" is much faster than running BLAST multiple times.

Modified from http://en.wikipedia.org/wiki/BLAST

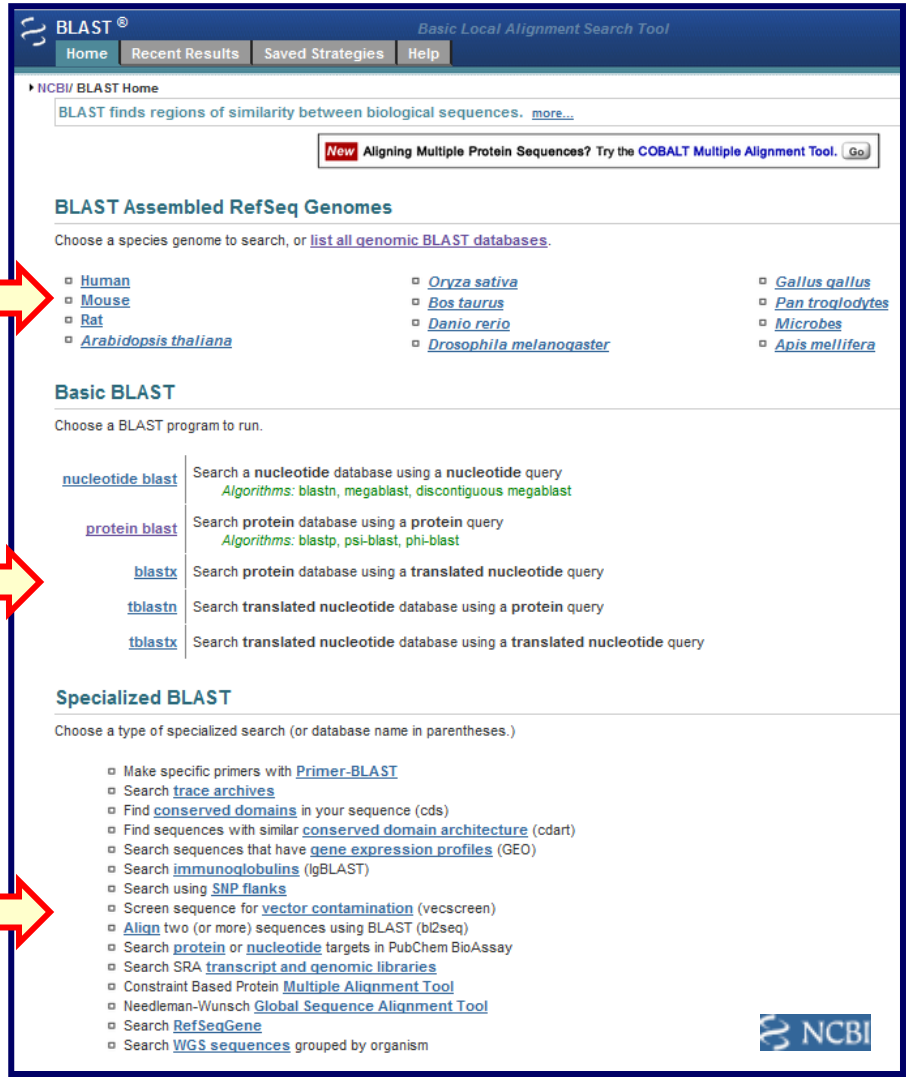
USING BLAST

The main page of BLAST has three primary areas:

- **BLAST Assembled Genomes** contains links to genomic BLAST pages for common organisms, and a link to a complete list of available organism genome BLAST pages. Here you can restrict your search to a particular species, such as human or mouse.

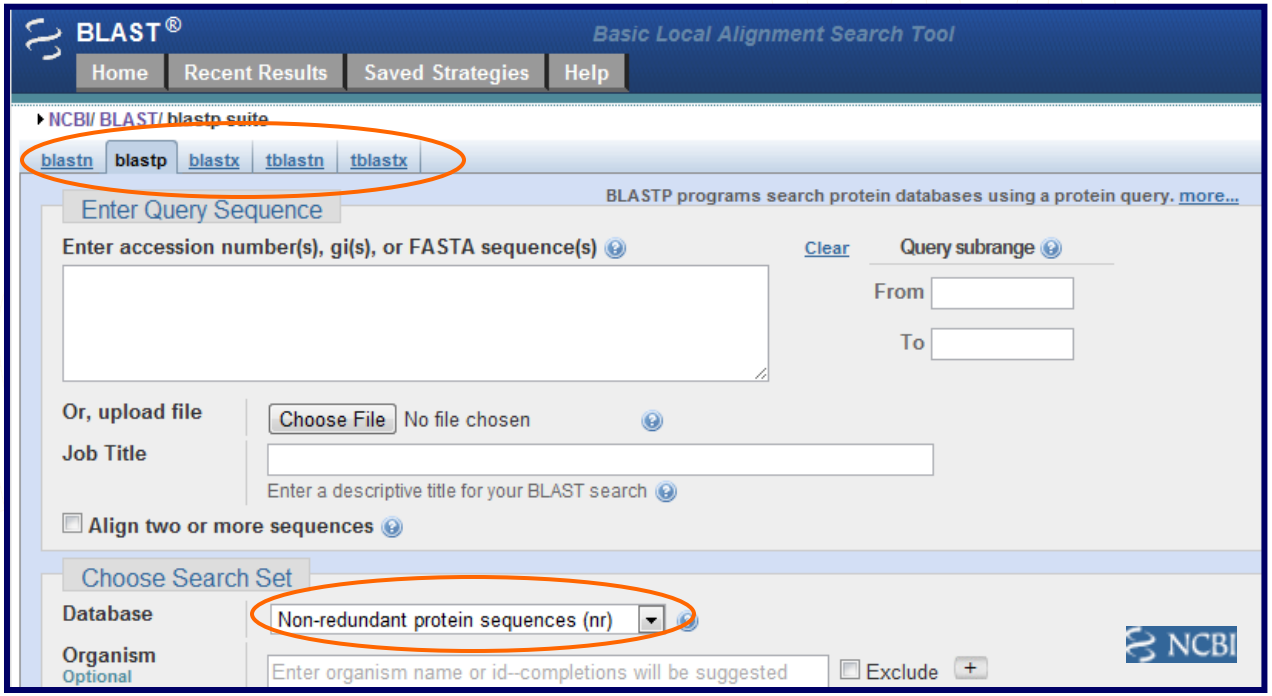
- **Basic BLAST** contains links to BLAST forms for the traditional set of databases (e.g., nr, est, etc.), such as nucleotide blast, protein blast, blastx, tblastx, etc.

- **Specialized BLAST** contains links to special-purpose BLAST databases and tools, e.g., looking for conserved domains, using optimal Needleman-Wunsch algorithm, multiple alignment tool, etc.



USING BLAST

- A BLAST search has four components:
  - Query: What is the sequence of interest?
  - Database: What database(s) would you like to search to find a possible alignment for the query sequence?
  - Program: Of the several variations of BLAST, which particular implementation are you interested in, e.g., blastn, blastx, tblastx, etc.
  - Search purpose/goal: What is the goal of the search? Is it finding a new gene, finding related genes, finding related species, determining the evolutionary distance between two genomes, etc.



BLAST PROTEIN DATABASES

- There are three groups of databases: Nucleotide databases; protein databases; specialized databases

DEFAULT

WHO?

Table 2.1 Content of Protein Sequence Databases

Database [1]	Content Description
nr	Non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF, excluding those in env_nr.
refseq	Protein sequences from NCBI Reference Sequence project.
swissprot	Last major release of the SWISS-PROT protein sequence database (no incremental updates).
pat	Proteins from the Patent division of GenBank.
month	All new or revised GenBank CDS translations + PDB + SwissProt + PIR + PRF released in the last 30 days.
pdb	Sequences derived from the 3-dimensional structure records from the Protein Data Bank.
env_nr	Non-redundant CDS translations from env_nt entries.
Smart v4.0 [2]	663 PSSMs from Smart, no longer actively maintained.
Pfam v11.0 [2]	7255 PSSMs from Pfam, not the latest.
COG v1.00 [2]	4873 PSSMs from NCBI COG set.
KOG v1.00 [2]	4825 PSSMs from NCBI KOG set (eukaryotic COG equivalent).
CDD v2.05 [2]	11399 PSSMs from NCBI curated cd set.

NOTE:
[1] default database is in bold.
[2] These databases are searchable only from rpsblast page, actual version may vary.

NCBI

BLAST NUCLEOTIDE DATABASES

DEFAULT

EASTERN STD. TIME?

Table 2.2 Nucleotide Databases for BLAST

Database	Content Description
nr [1]	All GenBank + EMBL + DDBJ + PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences). No longer "non-redundant" due to computational co[illegible]
refseq_mrna	[illegible]equences from NCBI Reference Sequence Project.
refseq_gen[illegible]	[illegible]omic sequences from NCBI Reference Sequence Project.
est	[illegible]atabase of GenBank + EMBL + DDBJ sequences from EST division.
est_human	Human subset of est.
est_mouse	Mouse subset of est.
est_others	Subset of est other than human or mouse.
gss	Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
htgs	Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2. Finished, phase 3 HTG sequences are in nr.
pat	Nucleotides from the Patent division of GenBank.
pdb	Sequences derived from the 3-dimensional structure records from Protein Data Bank. They are *NOT* the coding sequences for the coresponding proteins found in the same PDB record.
month	All new or revised GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days.
alu_repeats	Select Alu repeats from REPBASE, suitable for masking Alu repeats from query sequences. See "Alu alert" by Claverie and Makalowski, Nature 371: 752 (1994).
dbsts	Database of Sequence Tag Site entries from the STS division of GenBank + EMBL + DDBJ.
chromosome	Complete genomes and complete chromosomes from the NCBI Reference Sequence project. It overlaps with refseq_genomic.
wgs	Assemblies of **W**hole **G**enome **S**hotgun sequences.
env_nt	Sequences from environmental samples, such as uncultured bacterial samples isolated from soil or marine samples. The largest single source is Sagarsso Sea project. This does NOT overlap with nucleotide nr.

NOTE:
[1] default database is in bold.

NCBI

SPECIAL DATABASES
THE REFERENCE SEQUENCE PROJECT

- The ***Reference Sequence (RefSeq)*** is a special database built by the NCBI. It contains the nucleotide (DNA and RNA) and protein sequences; and unlike GenBank it provides a single, record for each naturally occurring molecule from a particular organism. Only major organisms are represented in the RefSeq database.
  - Each RefSeq is a synthesis of information, and not an actual research based data.
  - The collection explicitly maps the nucleotide and protein sequences: provides separate and linked records for the genomic DNA, the gene transcripts, and the proteins arising from those transcripts
- To produce RefSeq records, NCBI selects the best available information on each molecule and updates the records as more information emerges. A common analogy is that GenBank data come from primary research data, whereas RefSeq data is review of literature.
  - A RefSeq record may be just one really good example from GenBank copied to RefSeq, or generated by NCBI staff by combining several parts of GenBank records.
  - RefSeq records can be distinguished from GenBank records by their accession prefix, which includes an underscore, and a notation in the "comment" field that indicates the RefSeq status
- As of September 2011, the RefSeq Database includes 13^{+}million proteins from 16000^{+} organisms (as opposed to 250000^{+} organisms in GenBank)

SPECIAL DATABASES
THE REFERENCE SEQUENCE PROJECT

GENBANK	REFSEQ
Not curated	Curated
Author submits	NCBI creates from existing data
Only author can revise	NCBI revises as new data emerge
Multiple records for same loci common	Single records for each molecule of major organisms
Records can contradict each other	
No limit to species included	Limited to model organisms
Data exchanged among INSDC members	Exclusive NCBI database
Akin to primary literature	Akin to review articles
Proteins identified and linked	Proteins and transcripts identified and linked
Access via NCBI Nucleotide databases	Access via Nucleotide & Protein databases

http://www.ncbi.nlm.nih.gov/books/NBK21101/

SPECIAL DATABASES
EXPRESSED SEQUENCE TAGS (EST)

- An ***expressed sequence tag*** (EST) is a small portion of an entire gene, typically 200 – 500 nt long, used to help identify unknown genes and to map their positions within a genome. ESTs are so called because:
  - they are generated by sequencing either one or both ends of an *expressed* gene; and
  - they are short, used identify other genes, hence *tags*.
- An EST is obtained from a ***complimentary DNA*** (cDNA)
  - Recall that protein translation is done via mRNA, which travels from the nucleus to ribosomes, and only includes a copy of the exon (coding) regions of the DNA.
  - The problem is that mRNA is an unstable molecule once it is removed from the cell. To sequence it a lab, scientists use complimentary DNA (cDNA) instead, which is converted from mRNA using a special enzyme called ***reverse transcriptase***, so called because it is the reverse process of transcription.
  - cDNA is a much more stable molecule, and because it is obtained from the mRNA, it already has the introns removed, representing the expressed (protein coding) DNA sequence only.
- Once the cDNA is obtained
  - Sequencing only a few hundred nucleotides from each end (the 5' or the 3' end) creates the EST.
  - Now, not all of the gene's nucleotides are actually translated into a protein. The 5' end is what usually codes the protein, whereas the 3' end has untranslated regions (UTRs). It turns out the UTR exhibit less cross-species conservation than coding regions, hence *are usually unique* to each species.

SPECIAL DATABASES
EXPRESSED SEQUENCE TAGS (EST)

- How are ESTs used?
  - The 3' ESTs, because they are likely to be unique to each species, and because they point to an expressed gene, can serve as a "milemarker" on the genomic map of a particular species, precisely pointing to where a gene is located.
  - An EST, therefore, is an example of ***Sequenced Tag Site***, which is a short DNA sequence that is easily recognizable and occurs only once in a genome or a chromosome (hence a marker).
- Hence EST are used in gene mapping and gene finding
  - ESTs greatly reduce the time required to locate a gene, because the search can be done only on this short sequence, rather than the entire genome.
  - Because ESTs can be generated relatively easily, many ESTs are generated (about 70 million EST as of 2011), hence the necessity to create a separate database for them
- **dbEST** is that part of GenBank that includes the ESTs. There is also a separate database for the sequence tag sites, called the **dbSTS**.
  - NCBI scientists annotate the EST records with known information: If an EST matches a DNA sequence that codes a known gene with a known function, that gene's name and function are placed on the EST record.
  - Annotating EST records allows public scientists to use dbEST as an avenue for gene discovery.

WHICH BLAST PROGRAM FOR WHAT PURPOSE?

Nucleotide Queries

Length [1]	Database	Purpose	Program	Explanation
20 bp or longer 28 bp or above for megablast	Nucleotide	Identify the query sequence	discontiguous megablast, megablast, or blastn	Learn more ...
		Find sequences similar to query sequence	discontiguous megablast or blastn	Learn more ...
		Find similar sequence from the Trace archive	Trace megablast, or Trace discontiguous megablast	Learn more ...
		Find similar proteins to translated query in a translated database	Translated BLAST (tblastx)	Learn more ...
	Peptide	Find similar proteins to translated query in a protein database	Translated BLAST (blastx)	Learn more ...
7 - 20 bp	Nucleotide	Find primer binding sites or map short contiguous motifs	Search for short, nearly exact matches	Learn more ...

NOTE:
[1] The cut-off is only a recommendation. For short queries, one is more likely to get matches if the "Search for short, nearly exact matches" page is used. Detailed discussion is in the Section 4 below. With default setting, the shortest unambiguous query one can use is 11 for blastn and 28 for MEGABLAST. NCBI

WHICH BLAST PROGRAM FOR WHAT PURPOSE?

Protein Queries

Table 3.2 Program Selection for Protein Queries

Length [1]	Database	Purpose	Program	Explanation
15 residues or longer	Peptide	Identify the query sequence or find protein sequences similar to the query	Standard Protein BLAST (blastp)	Learn more ...
		Find members of a protein family or build a custom position-specific score matrix	PSI-BLAST	Learn more ...
		Find proteins similar to the query around a given pattern	PHI-BLAST	Learn more ...
		Find conserved domains in the query	CD-search (RPS-BLAST)	Learn more ...
		Find conserved domains in the query and identify other proteins with similar domain architectures	Conserved Domain Architecture Retrieval Tool (CDART)	Learn more ...
	Nucleotide	Find similar proteins in a translated nucleotide database	Translated BLAST (tblastn)	Learn more ...
5-15 residues	Peptide	Search for peptide motifs	Search for short, nearly exact matches	Learn more ...

Note:
[1] The cut-off is only a recommendation. For short queries, one is more likely to get matches if the "Search for short, nearly exact matches" page is used. Detailed discussion is in Section 4 below.

NCBI

WHICH BLAST PROGRAM FOR WHAT PURPOSE?

- Organism specific or genome databases

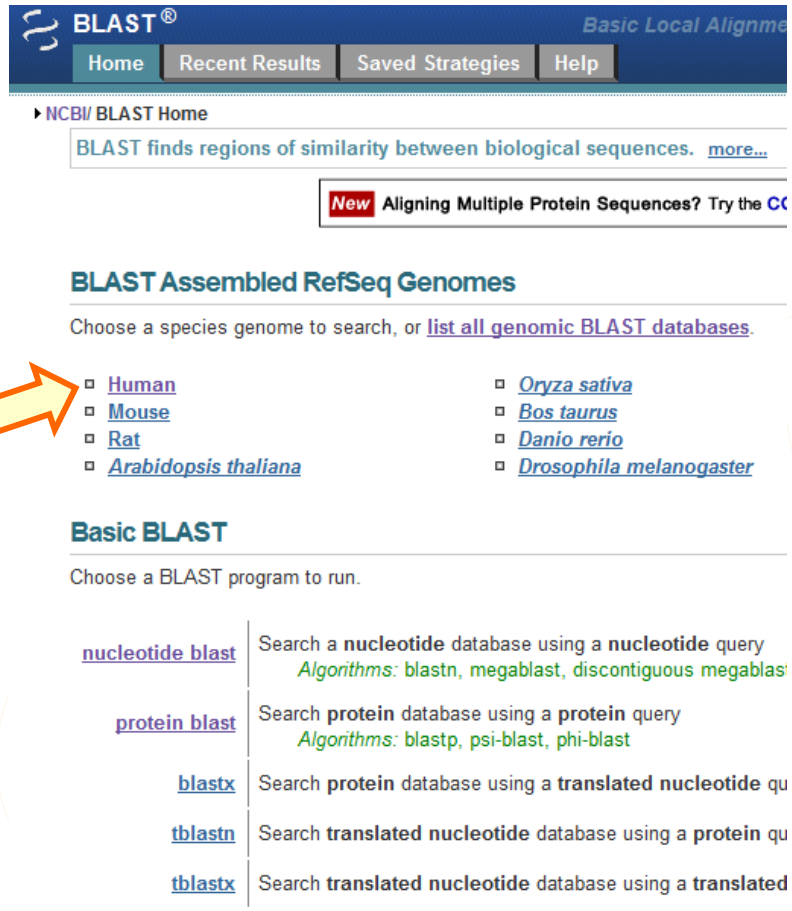
Query [2]	Database	Purpose	BLAST Pages to Use [3]	Explanation
Nucleotide: 20 or 28 bp and above Protein: 15 residues and above	Human Genome	Map the query sequence	Human	Learn more ...
	Mouse Genome		Mouse	Learn more ...
	Rat Genome		Rat	Learn more ...
	Chimp, Cow, Dog, or Chicken Genome		Chimp, or Cow, Dog, Chicken	Learn more ...
	Cat, Sheep, or Pig Genome		Cat, Sheep, or Pig	Learn more ...
	Zebrafish or Fugu (Pufferfish)	Determine the genomic structure	Zebrafish or Fugu rubripes	Learn more ...
	Insects (flies and honeybees)	Identify novel genes	Insects	Learn more ...
	Nematodes (worms)		Nematodes	Learn more ...
	Plants	Find homologs	Plants	Learn more ...
	Fungi Genomes (including yeasts)	Other data mining	Fungi	Learn more ...
	Protozoa		Protozoa	Learn more ...
	Environmental Samples		Environmental Samples	Learn more ...
	Other Lower Eukaryotic Genomes		Other eukaryotes genomes	Learn more ...
	Microbial Genomes		Microbial genomes	Learn more ...

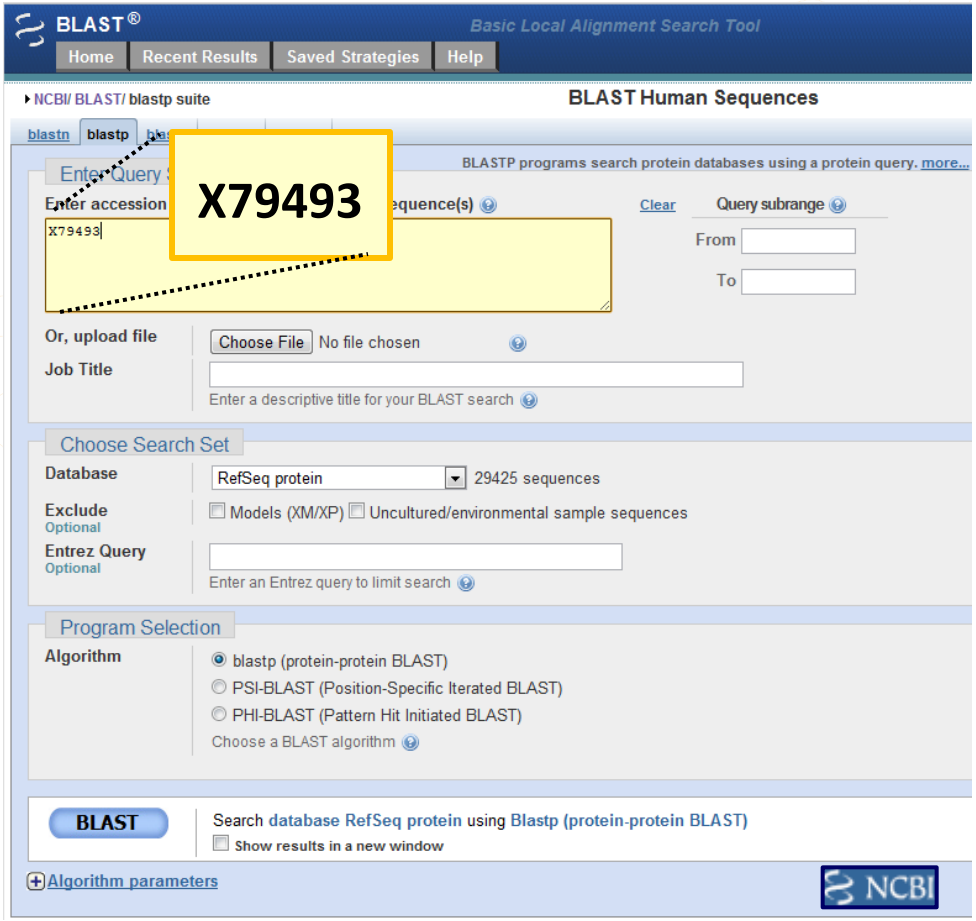
NOTE:
[1] Those pages access the genome database consisting of contig assemblies and other sequences specific to the organisms. Not all organisms listed here have genome assemblies available.
[2] Sequence length is only a suggestion. For most of the pages, the search parameters can be modified to enable searches with a short query by pasting additional options in the "Advanced Options" text box. For protein comparisons, -F F -e 20000 -W 2 should be used. For nucleotide comparison, use -F F -e 1000

NCBI

RUNNING BLAST

- You kind of need to know what you are searching. So, try the following:
  - Click on Human and type X79493, and hit BLAST!

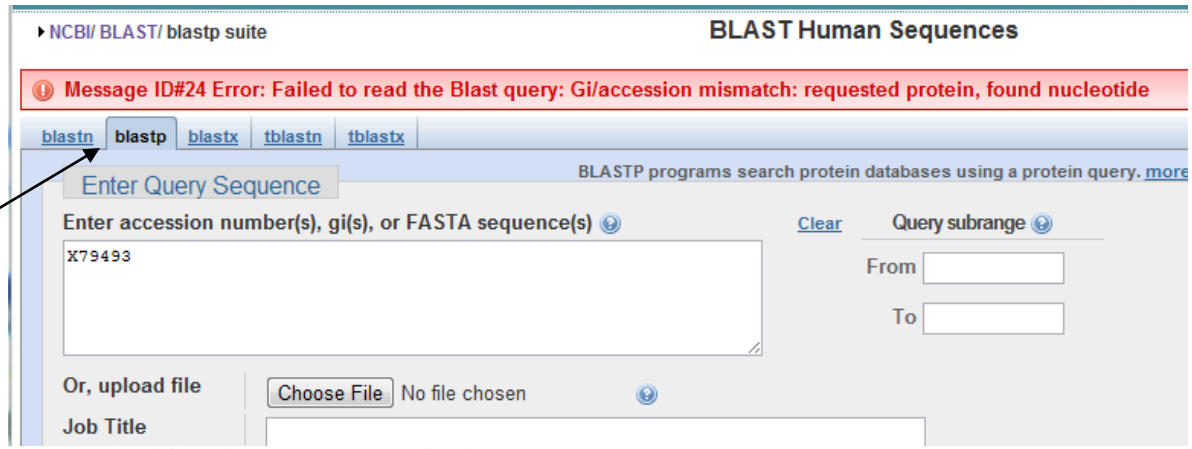


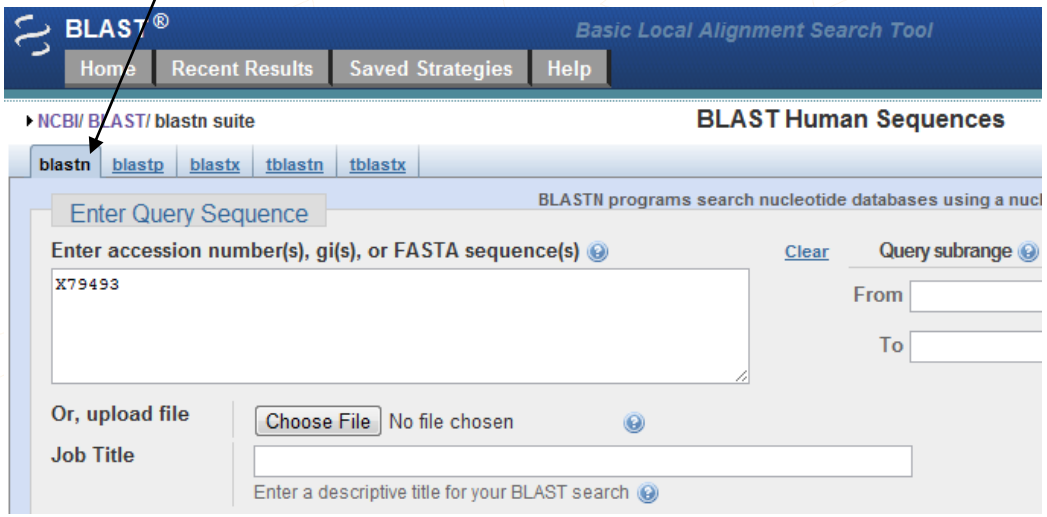


RUNNING BLAST

- Ohh…we have run a protein blast on a nucleotide sequence.
  - BLAST is smart enough to realize that is a nucleotide, but dumb enough not go ahead and run **blastn** on it…Well, you will have to do it.

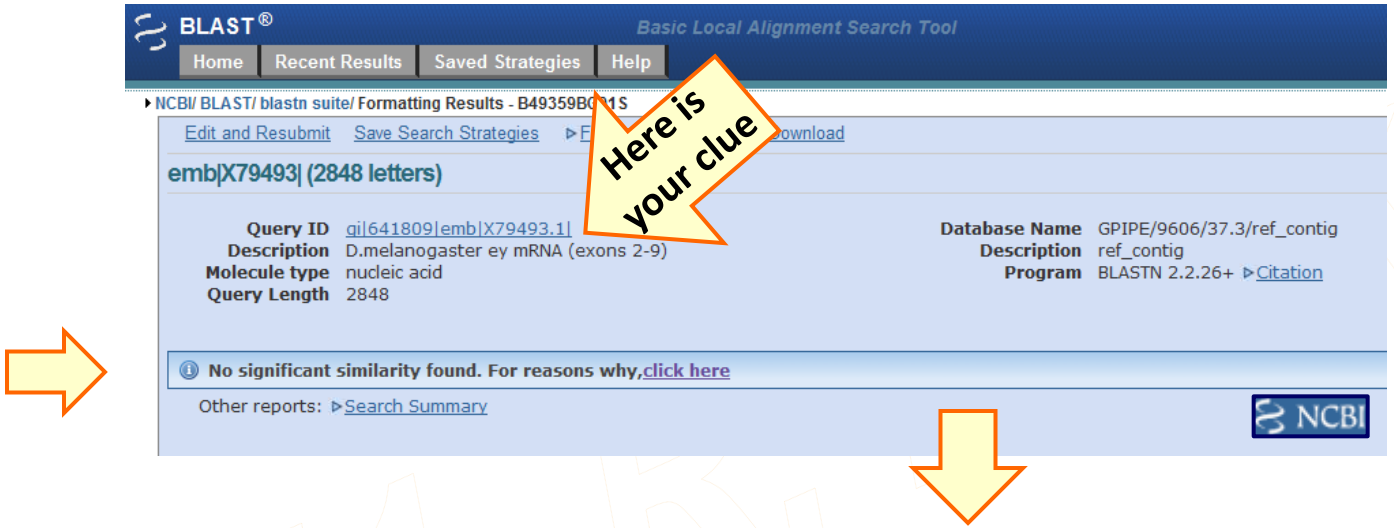
Choose the correct BLAST!

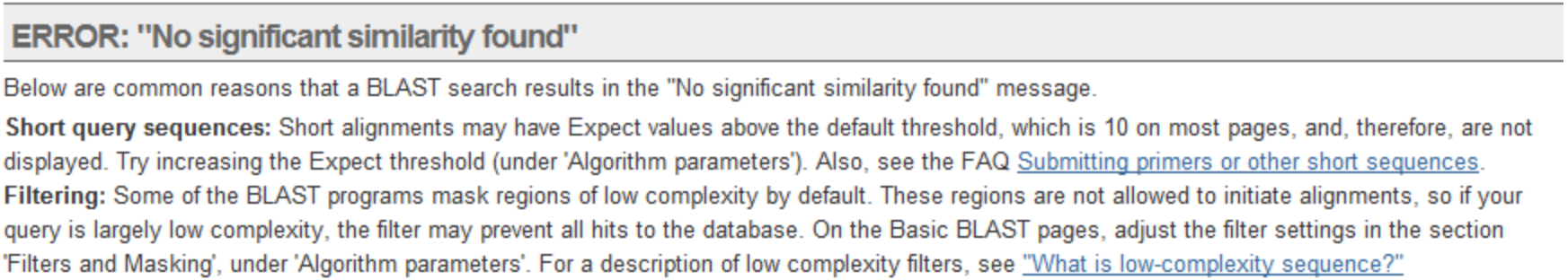




RUNNING BLAST

- What happened…? No significant similarity found…For reasons why, click here. Well, ok…

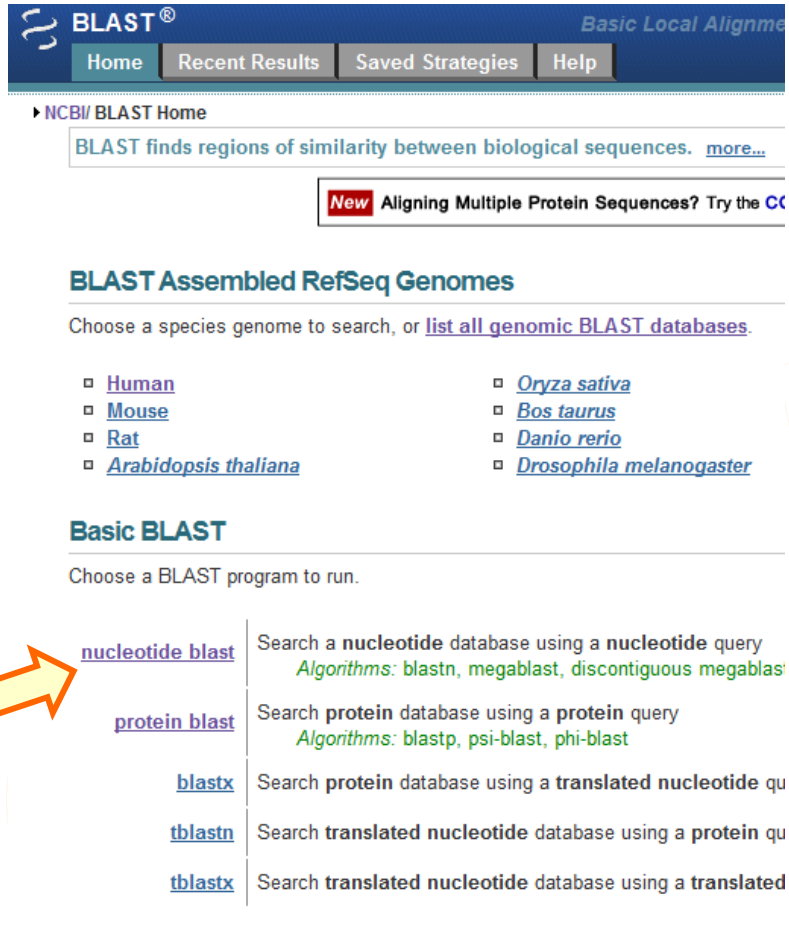


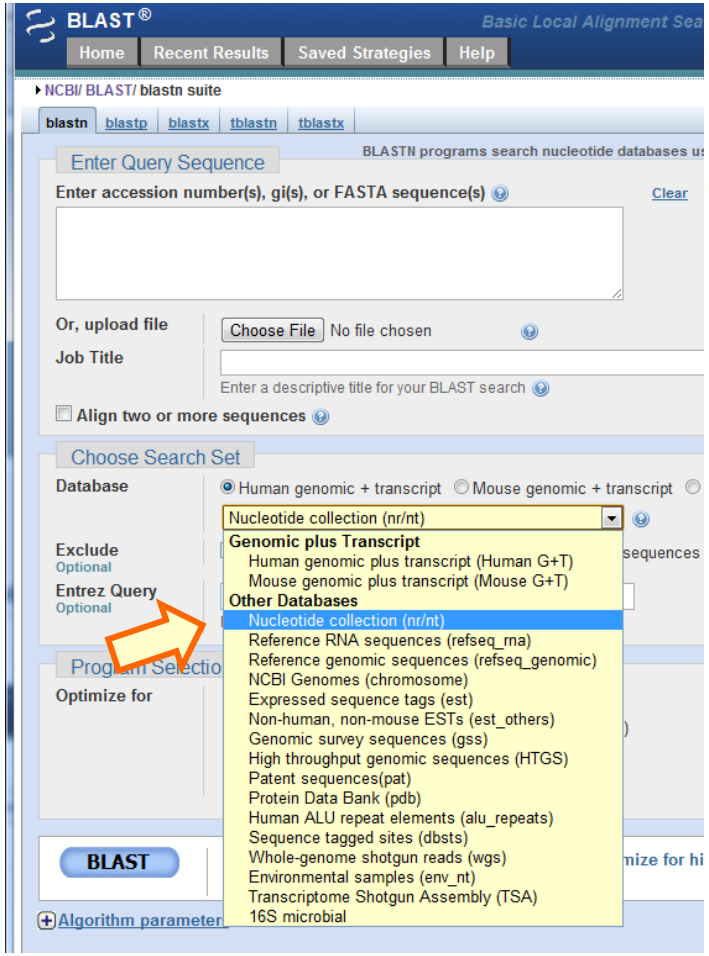


It turns out, it is none of these reasons! X79493 is a fruit fly…not a homosapien!

RUNNING BLAST

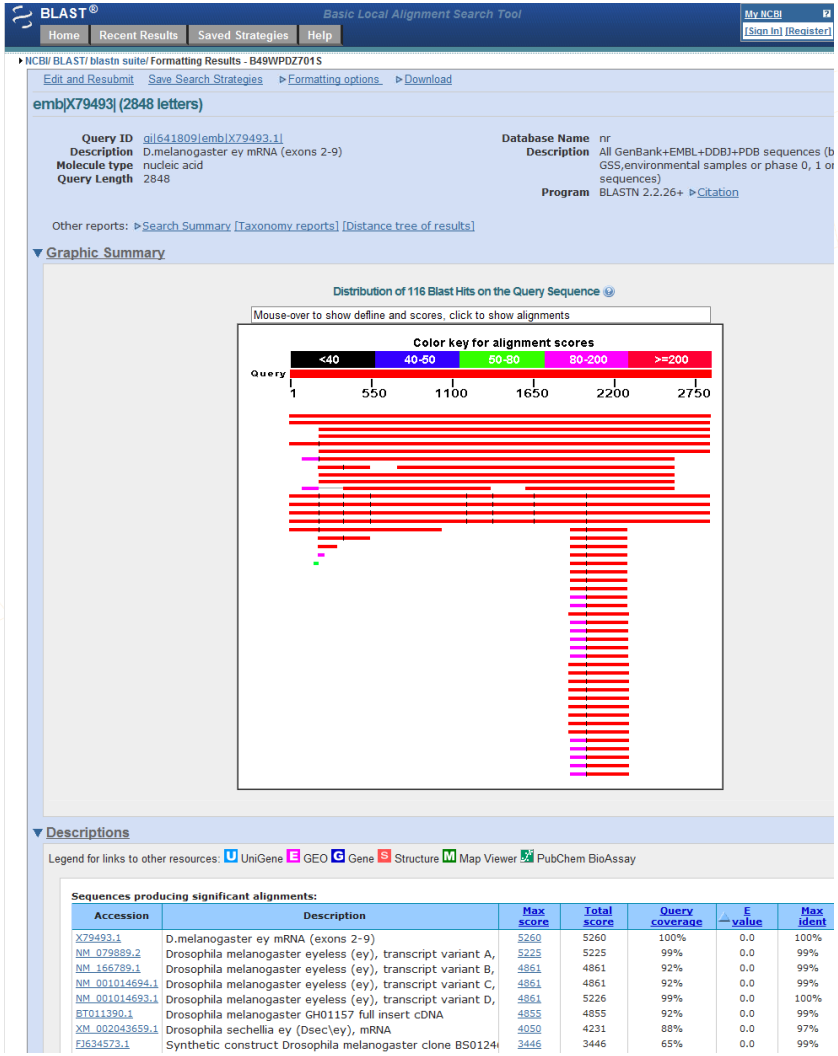
- Choose your BLAST programs and databases wisely!!!





RUNNING BLAST
SECTIONS OF BLAST REPORT

- Aha…some results…But what does it all mean?
- BLAST Results consists of five main regions
  - Header
  - Graphical summary
  - Description
  - Alignment



RUNNING BLAST
HEADER INFORMATION

- The Header includes such information as
  - Query ID
  - The name of the database that is searched
  - Description of the database that is searched
  - Description of the species detected (if any)
  - Query length
  - The type of BLAST that is used
- ...as well as links to
  - Search summary
  - Taxonomy reports
  - Distance tree results: TreeView, which disp ces matrix calculated from the BLAST local alignment r (more on this later in the semester)

Search Parameters	
Program	blastn
Word size	28
Expect value	10
Hitlist size	100
Match/Mismatch scores	1,-2
Gapcosts	0,0
Low Complexity Filter	Yes
Filter string	L;m;
Genetic Code	1

Database	
Posted date	Nov 1, 2011 4:13 PM
Number of letters	37,899,591,299
Number of sequences	14,874,993
Entrez query	none

Karlin-Altschul statistics		
Lambda	1.33271	1.28
K	0.620991	0.46
H	1.12409	0.85

Results Statistics	
Length adjustment	34
Effective length of query	2814
Effective length of database	37393841537
Effective search space	105226270085118
Effective search space used	105226270085118

NCBI/ BLAST/ blastn suite/ **Formatting Results - B49WPDZ701S**

Edit and Resubmit | Save Search Strategies | Formatting options | Downloa

emb|X79493| (2848 letters)

Query ID gi|641809|emb|X79493.1|
Description D.melanogaster ey mRNA (exons 2-9)
Molecule type nucleic acid
Query Length 2848

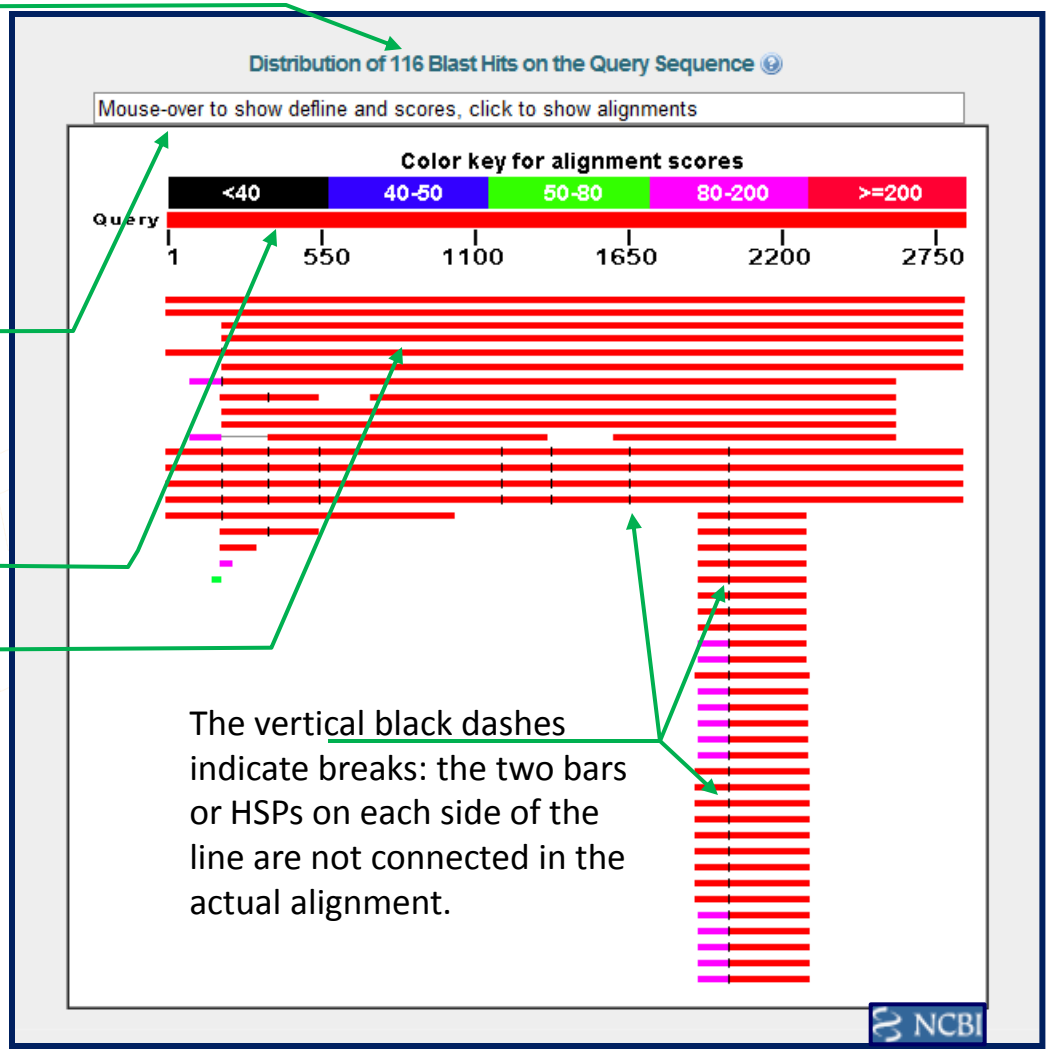
Other reports: Search Summary [Taxonomy reports] [Distance tree of res

EST, STS, GSS sequences)

NCBI

RUNNING BLAST
GRAPHICAL SUMMARY

- The number in the title is the total number of alignment segments (high scoring pairs, or HSPs), greater or equal to the setting in "Alignments" since some database sequences could have more than one segment aligned to the input query.
- The text box is for displaying the information on matching database sequences. Mouse-over the hits in the graph (colored bars), browser will display the information for that HSP in this box.
- Within the bordered graph, the top segment displays the color key and the query based scale. The colored bars represent the actual HSPs. The position of each bar indicates the region of the query the HSP covers.
- The bar color for a hit refers to alignment ***bit score S'***, a normalized value that reflects the degree of similarity between hit and query sequences.
  - The value S′ is derived from the raw alignment score S by normalizing w.r.t the scoring parameters K and λ. Because bit scores are normalized, they can be used to compare alignment scores from different searches.



Excellent match **Eh, Not bad Meh!** **You call this a match?** **Crap !**

RUNNING BLAST
DESCRIPTIONS

- This section provides quick one-line information about each of the alignments returned by BLAST. They are not complete descriptions (those come later), but provides an overview.
  - Each line has the following information: *the Accession number*; *the definition*: a brief textual description of the sequence that usually includes information on the organism from which the sequence was derived and the type of sequence (e.g., mRNA or DNA); *the alignment score* (bits) of that segment as well as the total score of that row; *query coverage* (how much of the query is aligned); *the E-value* as the statistical significance of the alignment (this is the default field by which the list is ordered); *max identity* (%), the maximum percentage of identical nucleotides or amino acids within the noted alignment length; Links to other databases, when the aligned target sequence is also part of a record from another database such as UniGene, GEO (Gene Expression Omnibus repository) profiles; Gene database, MapView , etc.

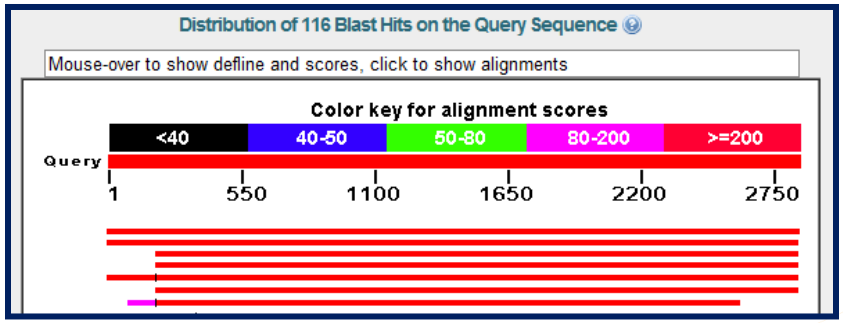
Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay NCBI

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
X79493.1	D.melanogaster ey mRNA (exons 2-9)	5260	5260	100%	0.0	100%	UM
NM_079889.2	Drosophila melanogaster eyeless (ey), transcript variant A, mRNA	5225	5225	99%	0.0	99%	UGM
NM_166789.1	Drosophila melanogaster eyeless (ey), transcript variant B, mRNA	4861	4861	92%	0.0	99%	UGM
NM_001014694.1	Drosophila melanogaster eyeless (ey), transcript variant C, mRNA	4861	4861	92%	0.0	99%	UGM
NM_001014693.1	Drosophila melanogaster eyeless (ey), transcript variant D, mRNA	4861	5226	99%	0.0	100%	UGM
BT011390.1	Drosophila melanogaster GH01157 full insert cDNA	4855	4855	92%	0.0	99%	UM
XM_002043659.1	Drosophila sechellia ey (Dsec\ey), mRNA	4050	4231	88%	0.0	97%	G
FJ634573.1	Synthetic construct Drosophila melanogaster clone BS01246 encc	3446	3446	65%	0.0	99%	
XM_001982674.1	Drosophila erecta GG16399 (Dere\GG16399), mRNA	3212	3212	84%	0.0	90%	G
XM_002099582.1	Drosophila yakuba GE14559 (Dyak\GE14559), mRNA	3049	3049	84%	0.0	89%	G
XM_002105728.1	Drosophila simulans eyeless (Dsim\ey), mRNA	1687	1872	38%	0.0	97%	G
XM_002105729.1	Drosophila simulans GD24412 (Dsim\GD24412), mRNA	1670	1670	35%	0.0	96%	G
AE014135.3	Drosophila melanogaster chromosome 4, complete sequence	1543	5268	99%	0.0	100%	E
AC150557.1	Drosophila melanogaster clone BACR06K04, complete sequence	1543	5268	99%	0.0	100%	
AC099309.1	Drosophila melanogaster, chromosome 4, region 101C-101D, BAC	1543	5268	99%	0.0	100%	
AC010576.16	Drosophila melanogaster, chromosome 4, region 101F-102F, BAC	1543	5268	99%	0.0	100%	
BT025949.2	Drosophila melanogaster IP14880 full insert cDNA	1531	1896	36%	0.0	100%	UM

RUNNING BLAST ALIGNMENTS

- Clicking on the actual colored bar in graphical display, or the Max Score in the description table, gives you the actual alignment for the selected target sequence.



Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score
X79493.1	D.melanogaster ey mRNA (exons 2-9)	5260
NM_079889.2	Drosophila melanogaster eyeless (ey), transcript variant A, mRNA	5225
NM_166789.1	Drosophila melanogaster eyeless (ey), transcript variant B, mRNA	4861
NM_001014694.1	Drosophila melanogaster eyeless (ey), transcript variant C, mRNA	4861
NM_001014693.1	Drosophila melanogaster eyeless (ey), transcript variant D, mRNA	4861

>emb|X79493.1| UM D.melanogaster ey mRNA (exons 2-9)
Length=2848

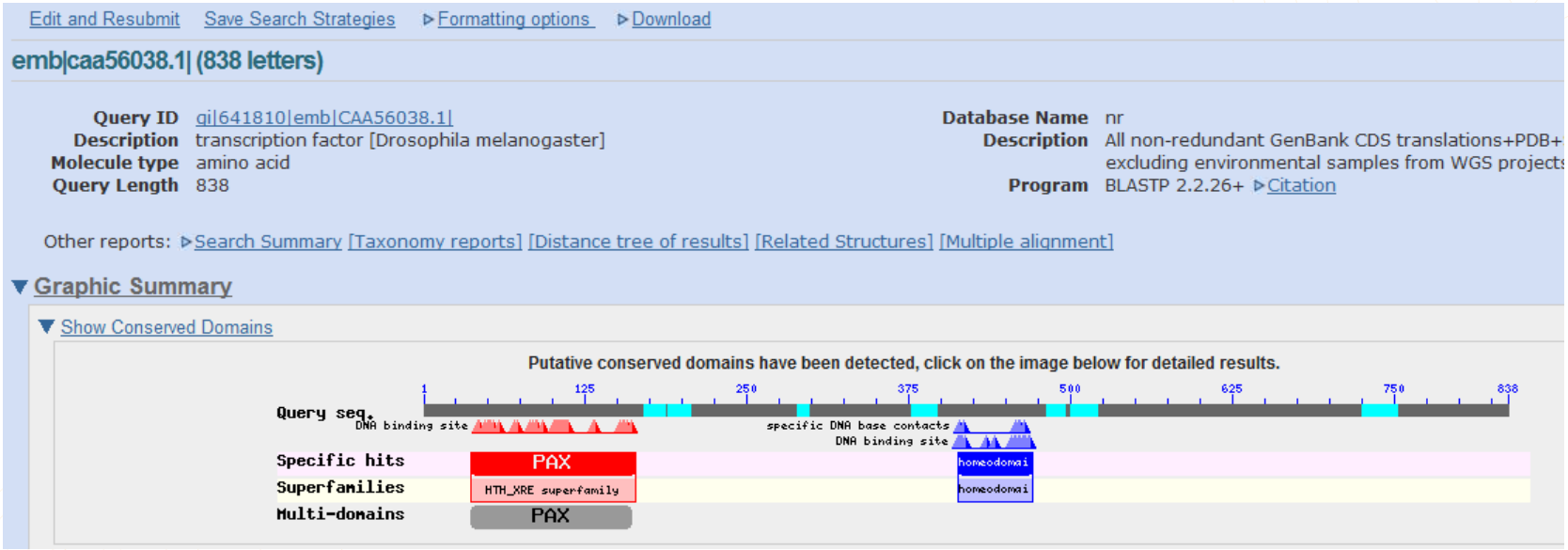
```
 Score = 5260 bits (2848),  Expect = 0.0
 Identities = 2848/2848 (100%), Gaps = 0/2848 (0%)
 Strand=Plus/Plus

Query  1     TTCGCACGGCGTGCGTTTGGCTGAACACAGCAGTCTCTTGGCTAAAGCTTTCATGAGCAG  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     TTCGCACGGCGTGCGTTTGGCTGAACACAGCAGTCTCTTGGCTAAAGCTTTCATGAGCAG  60

Query  61    TGCATGTAATAAAAACTGAGATCCAACTATGTTTACATTGCAACCAACTCCAACTGCTAT  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    TGCATGTAATAAAAACTGAGATCCAACTATGTTTACATTGCAACCAACTCCAACTGCTAT  120

Query  121   AGGCACCGTGGTTCCCCCATGGTCAGCGGGAACATTGATAGAGCGCCTGCCGTCTTTAGA  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   AGGCACCGTGGTTCCCCCATGGTCAGCGGGAACATTGATAGAGCGCCTGCCGTCTTTAGA  180

Query  181   AGACATGGCTCACAAGGGTCACAGTGGAGTAAATCAGCTGGGTGGCGTTTTTGTTGGAGG  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181   AGACATGGCTCACAAGGGTCACAGTGGAGTAAATCAGCTGGGTGGCGTTTTTGTTGGAGG  240
```

… which includes some basic statistics about the alignment, followed by the actual alignment itself.
The alignment (segment) on the left is a perfect match, with 100% of nucleotides matching, no gaps, no mismatches, and a perfect E-value of "0". Normally, you may see:

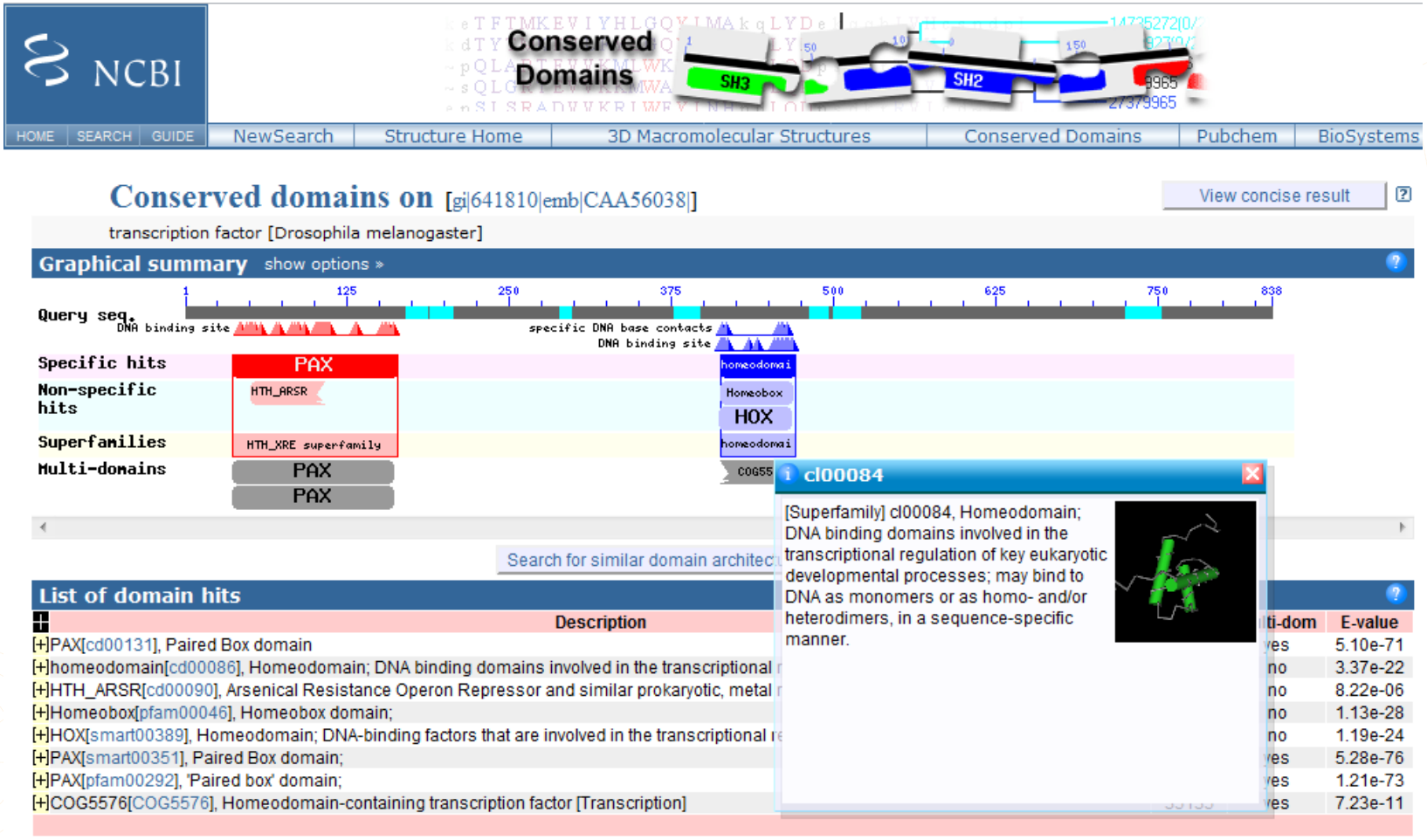
Pile (|): identity between query and database sequence
Space: mismatch dash
Dash(-): represents gap
Lowercase (gray): filter masked query (indicating low complexity and/or repeats)

RUNNING BLAST
PROTEIN SEQUENCES

- When using BLAST for protein sequences, such as with blastp, additional information can be returned by BLAST, such as conserved and other important domains:
  - Here is an example: CAA56038.1 is a transcription factor for the D.melanogaster, a 838AA long sequence



RUNNING BLAST
PROTEIN SEQUENCES



BLASTNCBI()

blastncbi Create remote NCBI BLAST report request ID or link to NCBI BLAST report

blastncbi(Seq, Program) sends a BLAST request to NCBI against a Seq, a nucleotide or amino acid sequence, using Program, a specified BLAST program, and then returns a command window link to the NCBI BLAST report. For help in selecting an appropriate BLAST program, visit: http://blast.ncbi.nlm.nih.gov/producttable.shtml
RID = **blastncbi**(Seq, Program) returns RID, the Request ID for the report.
[RID, RTOE] = **blastncbi**(Seq, Program) returns both RID, the Request ID for the NCBI BLAST report, and RTOE, the Request Time Of Execution, which is an estimate of the time until completion. **TIP:** You can use RTOE with the 'WaitTime' property when using the getblast function.
... **blastncbi**(Seq, Program, ...'Database', DatabaseValue, ...) specifies a database for the alignment search. Compatible databases depend on the type of sequence specified by Seq, and the program specified by Program. Database choices for nucleotide sequences are: 'nr' (default); 'refseq_rna'; 'refseq_genomic'; 'est'; 'est_human'; 'est_mouse'; 'est_others'; 'gss'; 'htgs'; 'pat'; 'pdb'; 'month'; 'alu_repeats'; 'dbsts'; 'chromosome'; 'wgs'; 'env_nt'. Database choices for amino acid sequences are: 'nr' (default); 'refseq_protein'; 'swissprot'; 'pat'; 'month'; 'pdb'; 'env_nr'. For help in selecting an appropriate database, visit: http://blast.ncbi.nlm.nih.gov/producttable.shtml
... **blastncbi**(Seq, Program, ...'Descriptions', DescriptionsValue, ...) specifies the number of short descriptions to include in the report, when you do not specify return values.
... **blastncbi**(Seq, Program, ...'Alignments', AlignmentsValue, ...) specifies the number of sequences for which high-scoring segment pairs (HSPs) are reported, when you do not specify return values. Default is 100.
... **blastncbi**(Seq, Program, ...'Filter', FilterValue, ...) specifies the filter to apply to the query sequence. Choices: 'L' Low (default); "R" Human repeats; "lcase": lower case mask
... **blastncbi**(Seq, Program, ...'Expect', ExpectValue, ...) specifies a statistical significance threshold for matches against database sequences. Choices are any real number. Default is 10. You can learn more about the statistics of local sequence comparison at: http://blast.ncbi.nlm.nih.gov/tutorial/Altschul-1.html#head2
... **blastncbi**(Seq, Program, ...'Word', WordValue, ...) specifies a word size for the query sequence. Choices for AA: 2, 3 (D); NT: 7, 11(D), 15
... **blastncbi**(Seq, Program, ...'Matrix', MatrixValue, ...) specifies the substitution matrix for amino acid sequences only. This matrix assigns the score for a possible alignment of two amino acid residues. Choices: PAM30, PAM70, BLOSUM45, BLOSUM62(D); BLOSUM(80)
... **blastncbi**(Seq, Program, ...'GapOpen', GapOpenValue, ...) specifies the penalty for opening a gap in the alignment of amino acid sequences. Choices and default depend on the substitution matrix specified by the 'Matrix' property.
... **blastncbi**(Seq, Program, ...'ExtendGap', ExtendGapValue, ...) specifies the penalty for extending a gap greater than one space in the alignment of amino acid sequences.
... **blastncbi**(Seq, Program, ...'GapCosts', GapCostsValue, ...) specifies the penalty for opening and extending a gap in the alignment of amino acid sequences. GapCostsValue is a vector containing two integers: the first is the penalty for opening a gap, and the second is the penalty for extending the gap.
... **blastncbi**(Seq, Program, ...'Inclusion', InclusionValue, ...) specifies the statistical significance threshold for including a sequence in the Position-Specific Scoring Matrix (PSSM) created by PSI-BLAST for the subsequent iteration. Default is 0.005. **Note :** Specify an InclusionValue only when Program = 'psiblast'.
... **blastncbi**(Seq, Program, ...'Pct', PctValue, ...) specifies the percent identity and the corresponding match and mismatch score for matching existing sequences in a public database. Default is 99. **Note :** Specify a PctValue only when Program = 'megablast'.
... **blastncbi**(Seq, Program, ...'Entrez', EntrezValue, ...) specifies Entrez query syntax to search a subset of the selected database.
Note For more information about Entrez query syntax, see: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez.chapter.EntrezHelp

BLASTNCBI()
OPTIONAL PARAMETERS

Choices for Optional Properties by BLAST Program

When BLAST program is...	Then choices for the following properties are...		
	Database	Filter	Word
'blastn'	'nr' (default) 'est' 'est_human' 'est_mouse' 'est_others' 'gss' 'htgs' 'pat' 'pdb' 'month' 'alu_repeats' 'dbsts' 'chromosome' 'wgs' 'refseq_rna' 'refseq_genomic' 'env_nt'	'L' (default) 'R' 'm' 'lcase'	7 11 (default) 15
'megablast'		'L'	11 12 16 20 24 28 (default) 32 48 64
'tblastn'		'L' (default) 'm' 'lcase'	2 3 (default)
'tblastx'		'L' (default) 'R' 'm' 'lcase'	
'blastp'	'nr' (default) 'swissprot' 'pat' 'pdb' 'month' 'refseq_protein' 'env_nr'	'L' (default) 'm' 'lcase'	
'blastx'			
'psiblast'			

Choices for the GapCosts Property by Matrix

When substitution matrix is...	Then choices for GapCosts are...
'PAM30'	[7 2] [6 2] [5 2] [10 1] [9 1] (default) [8 1]
'PAM70'	[8 2] [7 2] [6 2] [11 1] [10 1] (default) [9 1]
'BLOSUM80'	
'BLOSUM45'	[13 3] [12 3] [11 3] [10 3] [15 2] (default) [14 2] [13 2] [12 2] [19 1] [18 1] [17 1] [16 1]
'BLOSUM62'	[9 2] [8 2] [7 2] [12 1] [11 1] (default) [10 1]

SAMPLE USAGE

```
RID = blastncbi('AAA59174','blastp','matrix','PAM70', 'expect',1e-10)
Struct = getblast(RID)
```

```
S = getpdb('1CIV')
blastncbi(S,'blastp','expect',1e-10)
```

OTHER RELATED MATLAB FUNCTION

getblast() : Retrieve BLAST report from NCBI Web site

Data = **getblast**(RID) reads RID, the Request ID for the NCBI BLAST report, and returns the report data in Data, a MATLAB structure or array of structures. The Request ID, RID, must be recently generated because NCBI purges reports after 24 hours. The function returns a rich set of information about the requested search.

```
RID = blastncbi('AAA59174','blastp','expect',1e-10);
report = getblast(RID,'ToFile','AAA59174_BLAST.rpt')
        RID: '1175093633-2786-174709873694.BLASTQ3'
        Algorithm: 'BLASTP 2.2.16 [Mar-11-2007]'
        Query: [1x63 char]
        Database: [1x96 char]
        Hits: [1x50 struct]
        Statistics: [1x1034 char]
```

BLASTLOCAL()

blastlocal(): Perform search on local BLAST database to create BLAST report

This function assumes that BLAST offers a fast and powerful comparative analysis of protein and nucleotide sequences against known sequences in online or local databases. In order to use the **blastlocal** function, you must have a local copy of the NCBI blastall executable file (version 2.2.17) available from your system. You can download the blastall executable file by accessing http://blast.ncbi.nlm.nih.gov/download.shtml

blastlocal('InputQuery', InputQueryValue) submits query sequence(s) specified by InputQueryValue, a FASTA file containing nucleotide or amino acid sequence(s), for a BLAST search of a local BLAST database, by calling a local version of the NCBI blastall executable file. The BLAST search results are displayed in the MATLAB Command Window. (This corresponds to the blastall option -i.)

Data = **blastlocal**('InputQuery', InputQueryValue) returns the BLAST search results in Data, a MATLAB structure or array of structures (if multiple query sequences) containing fields corresponding to BLAST keywords and data from a local BLAST report.

LAB 2

1. Explore the X79493 and AY707088 genes:
   - Provide some background regarding these genes
   - Study their local and global alignments using Matlab tools (dotplot, NW, SW algorithms)
   - Study their local and global alignments using BLAST tools
   - Compare and interpret your results. Make sure to include
     - Statistical significance of these results.
     - Biological significance of these results
   - Repeat the above using protein sequences of the above two.
2. Find two other homologs, preferably highly conserved among species, and repeat Exercise 1.

PROJECT IDEAS

- Using a small subset dataset (of say 100-200 sequences)
  - Implement FASTA in Matlab.
  - Implement BLAST in Matlab.
    - For either of these two, compare your implementation to one found in FASTA / BLAST servers. I do not expect your implementation to be faster, but at least find the same alignment.

- Implement NW or SW align procedures in Matlab. Compare your implementation to the Matlab implementation.
  - I do not expect your implementation to be faster, but at least find the same alignment.

- More coming soon.