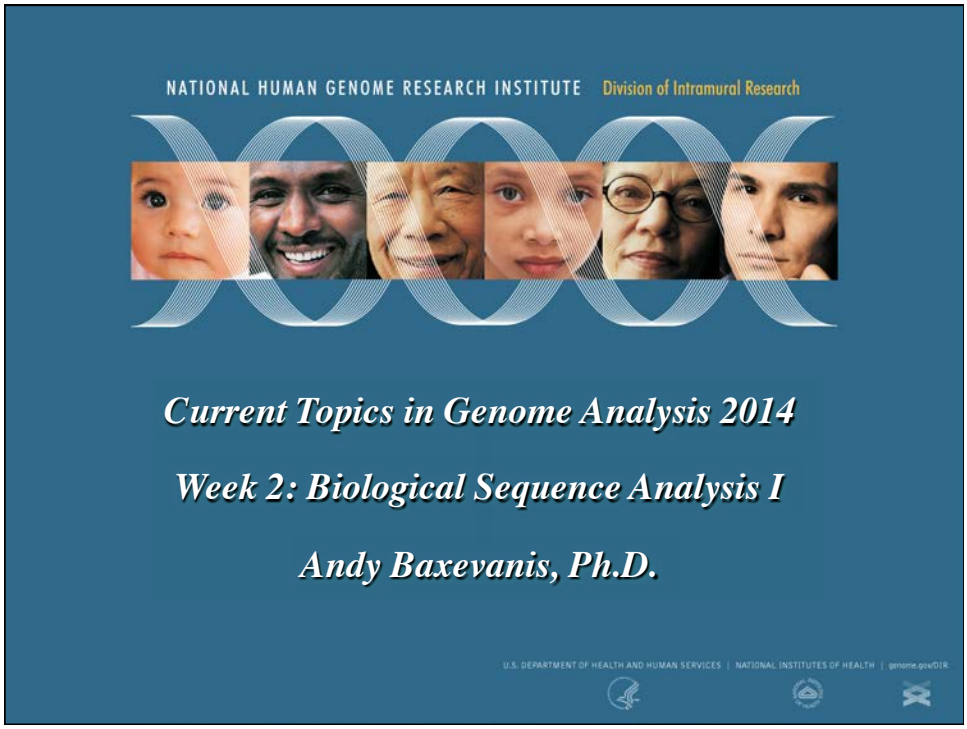
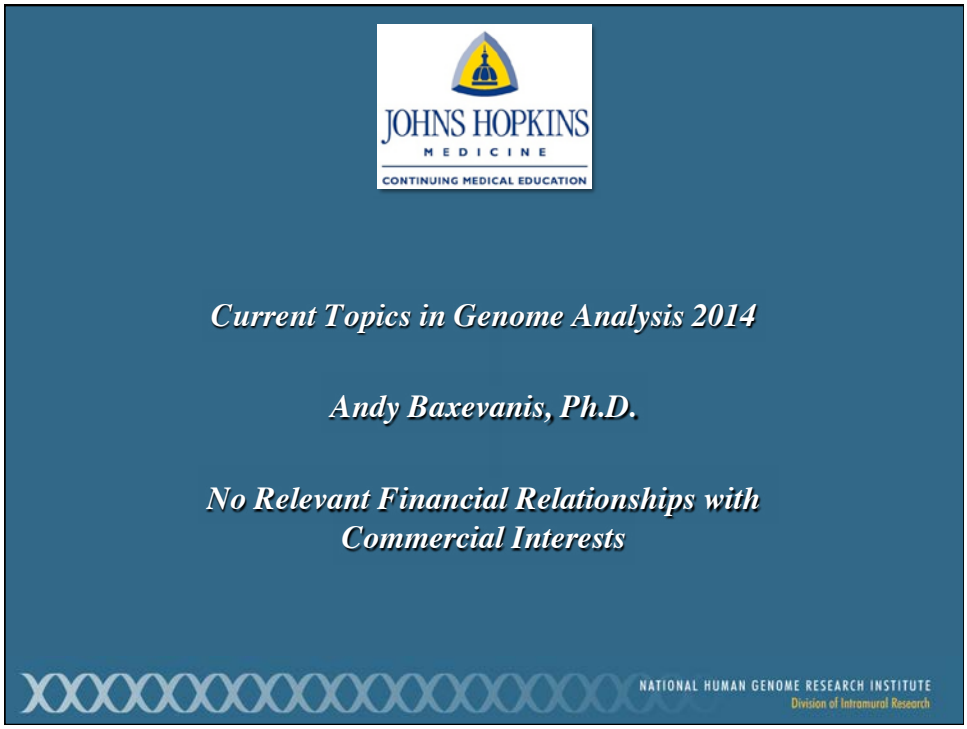
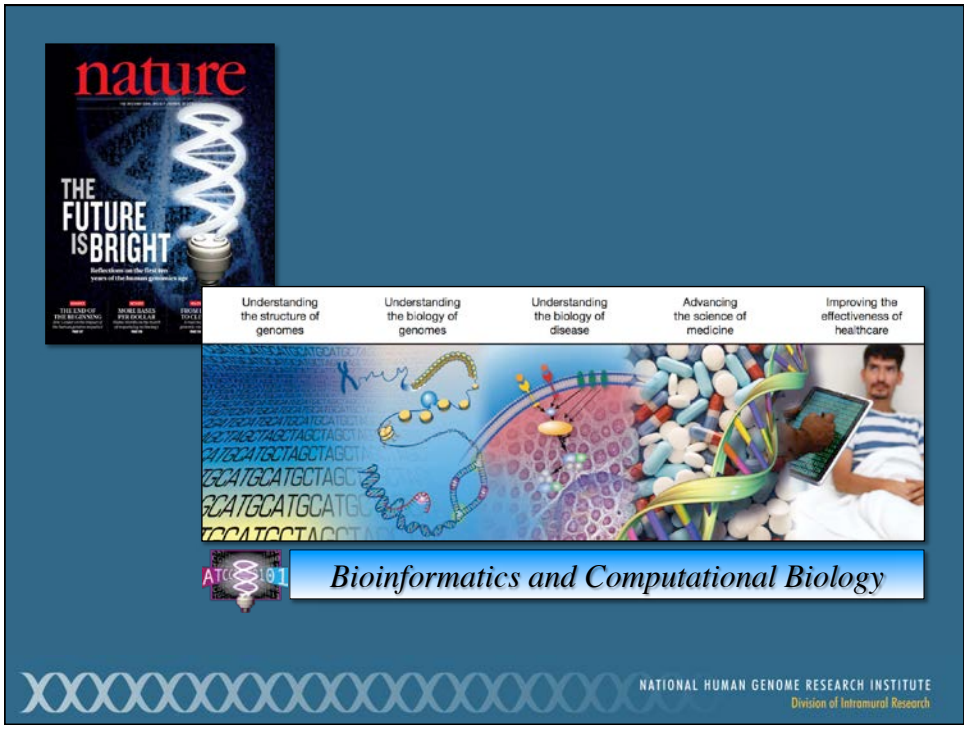
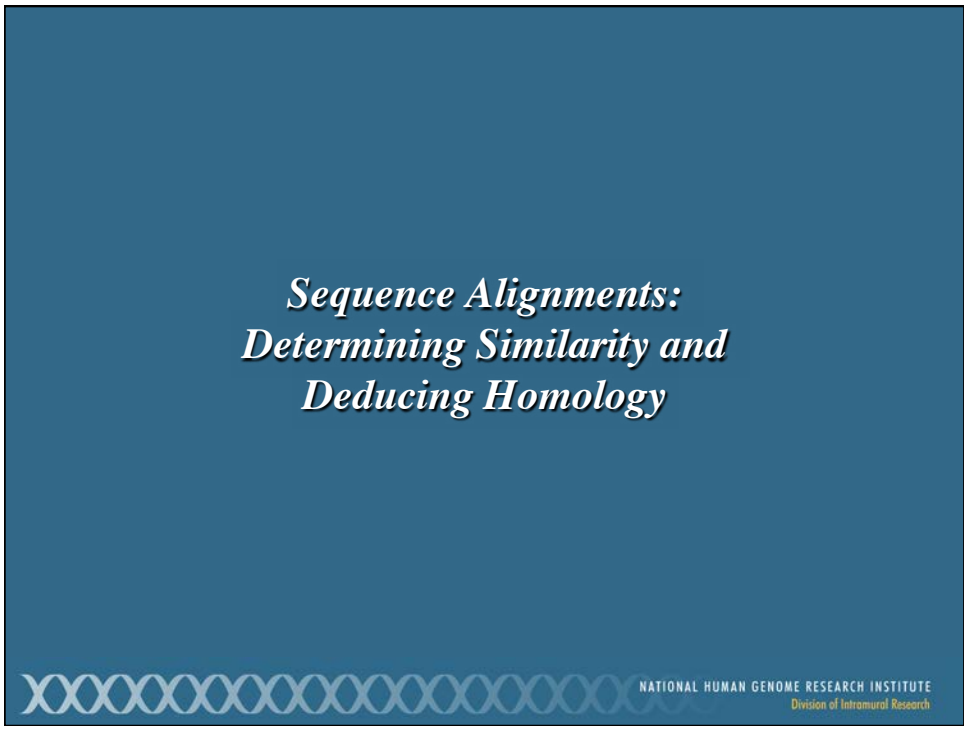

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research
Current Topics in Genome Analysis 2014
Week 2: Biological Sequence Analysis I
Andy Baxevanis, Ph.D.
U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES | NATIONAL INSTITUTES OF HEALTH | genome.gov/DIR


JOHNS HOPKINS
MEDICINE
CONTINUING MEDICAL EDUCATION
Current Topics in Genome Analysis 2014
Andy Baxevanis, Ph.D.
No Relevant Financial Relationships with Commercial Interests
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research


nature
THE FUTURE IS BRIGHT
Understanding the structure of genomes
Understanding the biology of genomes
Understanding the biology of disease
Advancing the science of medicine
Improving the effectiveness of healthcare
Bioinformatics and Computational Biology
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research


Sequence Alignments:
Determining Similarity and
Deducing Homology
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Why construct sequence alignments?

- Provide a measure of relatedness between nucleotide or amino acid sequences
- Determining relatedness allows one to draw biological inferences regarding
  - structural relationships
  - functional relationships
  - evolutionary relationships
- Important to use correct terminology when describing phylogenetic relationships

Defining the Terms

- The quantitative measure: ***Similarity***
  - Always based on an observable
  - Usually expressed as percent identity
  - Quantify changes that occur as two sequences diverge (substitutions, insertions, or deletions)
  - Identify residues crucial for maintaining a protein's structure or function
- High degrees of sequence similarity *might* imply
  - a common evolutionary history
  - possible commonality in biological function

Defining the Terms

The conclusion: ***Homology***

- ***Homology***: Implies an evolutionary relationship
- ***Homologs***: Genes that have arisen from a common ancestor
- Genes either *are* or *are not* homologous
  (not measured in degrees)

> It is worth repeating here that homology, like pregnancy, is indivisible[8]. You either are homologous (pregnant) or you are not. Thus, if what one means to assert is that 80% of the character states are identical one should speak of 80% identity, and not 80% homology.

Fitch, Trends Genet. 16: 227-231, 2000

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Defining the Terms

Orthologs: Genes that diverged as a result of a speciation event

- Sequences are direct descendants of a sequence in a common ancestor
- Most likely have similar domain and three-dimensional structure
- Usually retain same biological function over evolutionary time
- Can be used to predict gene function in novel genomes

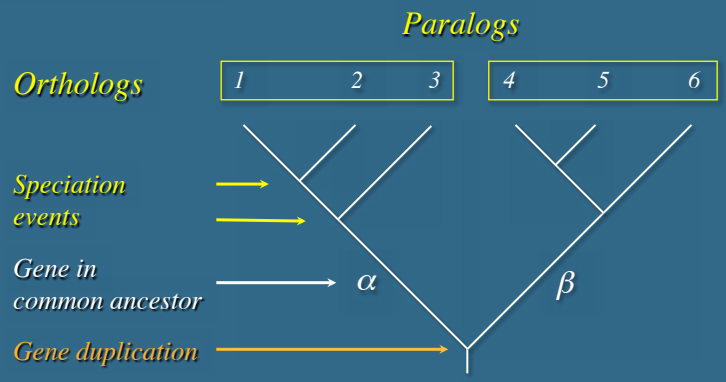
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Defining the Terms

Paralogs: Genes that arose by the duplication of a single gene in a particular lineage

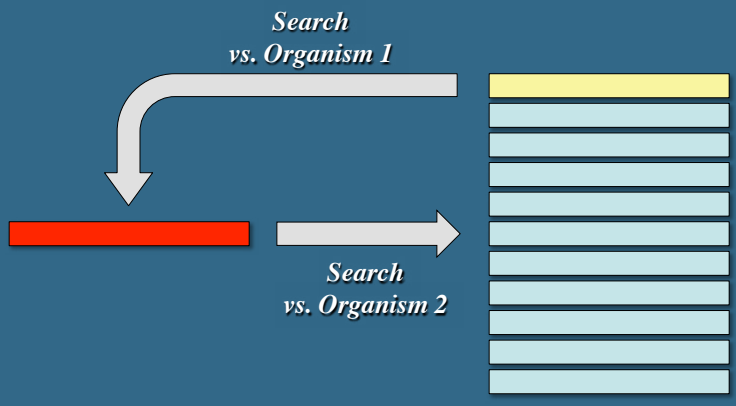
- Perhaps less likely to perform similar functions
- Can take on new functions over evolutionary time
- Provides insight into "evolutionary innovation"

Defining the Terms



- Genes 1-3 are orthologous
- Genes 4-6 are orthologous
- Any pair of α and β genes are paralogous (genes related through a gene duplication event)

Similarity Searching to Identify Reciprocal Best Hits



Global Sequence Alignments

- Sequence comparison along the entire length of the two sequences being aligned
- Best for highly-similar sequences of similar length
- As the degree of sequence similarity declines, global alignment methods tend to miss important biological relationships

Local Sequence Alignments

- Sequence comparison intended to find the most similar regions in the two sequences being aligned ("paired subsequences")
- Regions outside the area of local alignment are excluded
- More than one local alignment could be generated for any two sequences being compared
- Best for sequences that share some similarity, or for sequences of different lengths

Scoring Matrices: Construction and Proper Selection

Scoring Matrices

- Empirical weighting scheme representing physicochemical and biological characteristics of nucleotides and amino acids
  - Side chain structure and chemistry
  - Side chain function
- Amino acid-based examples of considerations:
  - Cys/Pro important for structure and function
  - Trp has bulky side chain
  - Lys/Arg have positively charged side chains

Scoring Matrices

- ***Conservation:*** What residues can substitute for another residue and not adversely affect the function of the protein?
  - Ile/Val - both small and hydrophobic
  - Ser/Thr - both polar
  - *Conserve charge, size, hydrophobicity, additional physicochemical factors*
- ***Frequency:*** How often does a particular residue occur amongst the entire constellation of proteins?

Scoring Matrices

Why is understanding scoring matrices important?

- Appear in all analyses involving sequence comparison
- Implicitly represent particular evolutionary patterns
- Choice of matrix can strongly influence outcomes of analyses

Matrix Structure: Nucleotides

- Simple match/mismatch scoring scheme:

  Match +2
  Mismatch −3

- Assumes each nucleotide occurs 25% of the time

	A	T	G	C
A	2	−3	−3	−3
T	−3	2	−3	−3
G	−3	−3	2	−3
C	−3	−3	−3	2

Matrix Structure: Proteins

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

BLOSUM62

BLOSUM Matrices

- Look only for differences in conserved, ungapped regions of a protein family ("blocks")
- Directly calculated based on local alignments
  - Substitution probabilities (*conservation*)
  - Overall *frequency* of amino acids
- Sensitive to detecting structural or functional substitutions
- Generally perform better than PAM matrices for local similarity searches *(Henikoff and Henikoff, 1993)*
- BLOSUM series can be used to identify both closely and distantly related sequences

BLOSUM *n*

- Built using sequences sharing no more than *n*% identity
- Contribution of sequences > *n*% identical clustered and replaced by a sequence that represents the cluster

BLOSUM *n*

- Clustering reduces contribution of closely related sequences (less bias towards substitutions that occur in the most closely related members of a family)
- Reducing *n* yields more distantly related sequences
- Increasing *n* yields more closely related sequences

Which one to choose?

BLOSUM		% Similarity
90	Short alignments, highly similar	70-90
80	Best for detecting known members of a protein family	50-60
62	Most effective in finding all potential similarities	30-40
30	Longer, weaker local alignments	< 30

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

So many matrices...

No single matrix is the complete answer for all sequence comparisons

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Gaps

- Used to improve alignments between two sequences
  - Compensate for insertions and deletions
  - As such, *gaps represent biological events*
- Must be kept to a reasonable number, to not reflect a biologically implausible scenario (~1 gap per 20 residues good rule-of-thumb)
- Cannot be scored simply as a "match" or a "mismatch"

Affine Gap Penalty

Fixed deduction for introducing a gap *plus*
an additional deduction proportional to the length of the gap

Deduction for a gap = $G + Ln$

			nucleotide	protein
where	G =	gap-opening penalty	5	11
	L =	gap-extension penalty	2	1
	n =	length of the gap		
and	$G > L$			

BLAST:
The Basic Local Alignment Search Tool

BLAST

- Seeks high-scoring segment pairs (HSPs)
  - Pair of sequences that can be aligned with one another
  - When aligned, have maximal aggregate score (score cannot be improved by extension or trimming)
  - Score must be above score threshhold *S*
  - Gapped or ungapped
- Results not limited to the "best" high-scoring segment pair for the two sequences being aligned

Altschul et al., J. Mol. Biol. 215: 403-410, 1990

BLAST Algorithms

Program	*Query Sequence*	*Target Sequence*
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nucleotide, six-frame translation	Protein
TBLASTN	Protein	Nucleotide, six-frame translation
TBLASTX	Nucleotide, six-frame translation	Nucleotide, six-frame translation

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Neighborhood Words

Query Word (W = 3)

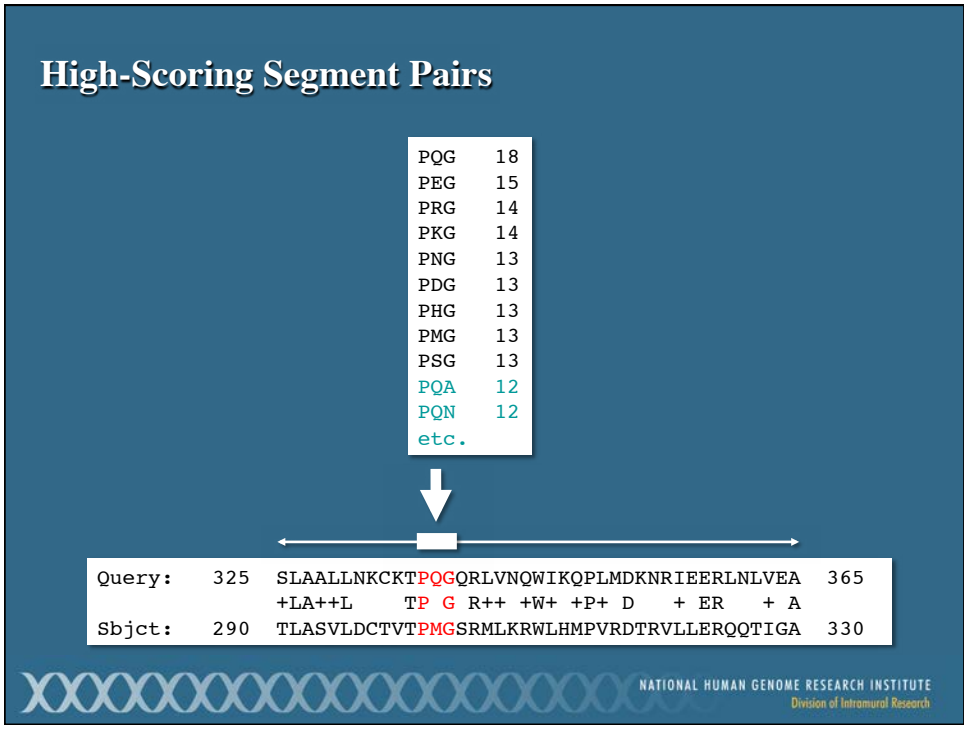
```
Query:   GSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVED
```

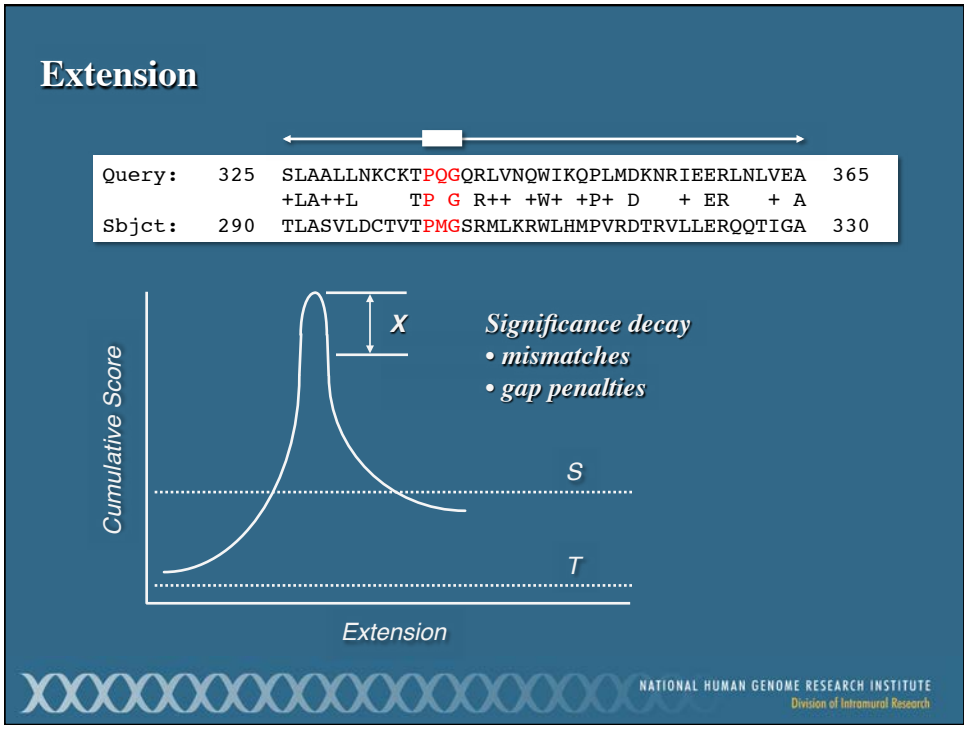
Neighborhood Words

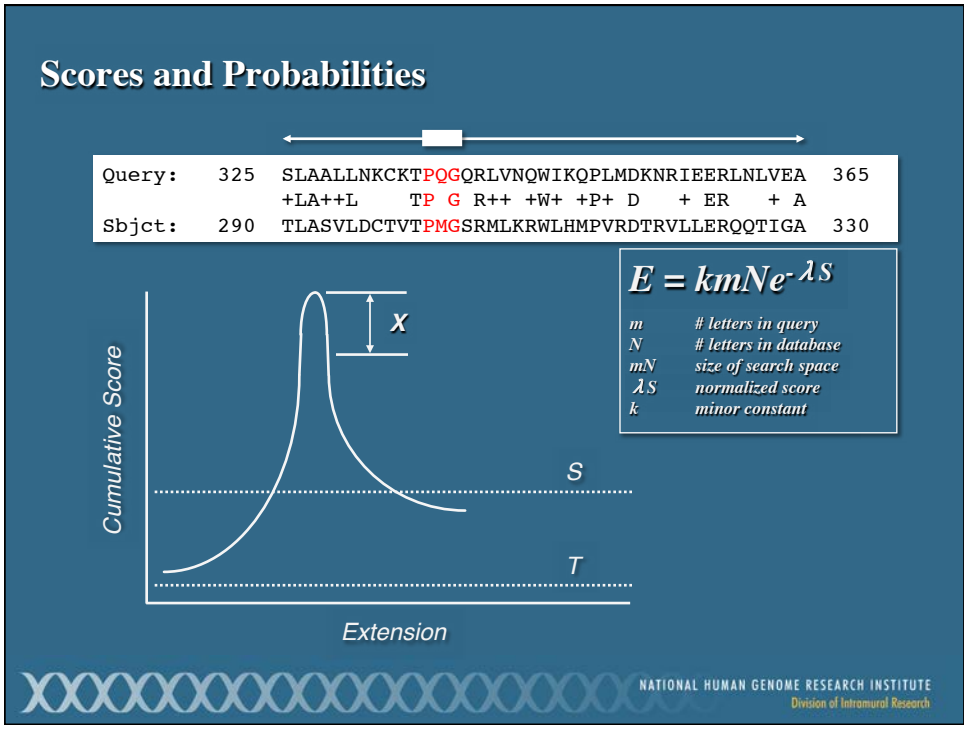
PQG	18	= 7 + 5 + 6
PEG	15	
PRG	14	
PKG	14	
PNG	13	
PDG	13	
PHG	13	
PMG	13	
PSG	13	
PQA	12	
PQN	12	
etc.		

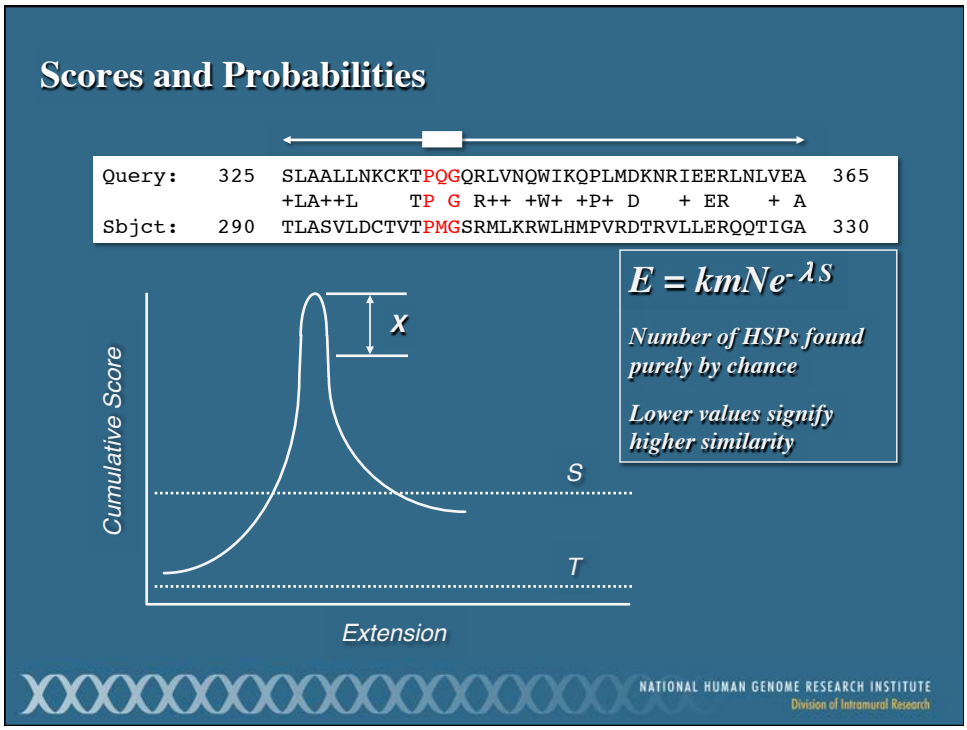
Neighborhood Score Threshold (T = 13)

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research


High-Scoring Segment Pairs
PQG 18
PEG 15
PRG 14
PKG 14
PNG 13
PDG 13
PHG 13
PMG 13
PSG 13
PQA 12
PQN 12
etc.
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L TP G R++ +W+ +P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

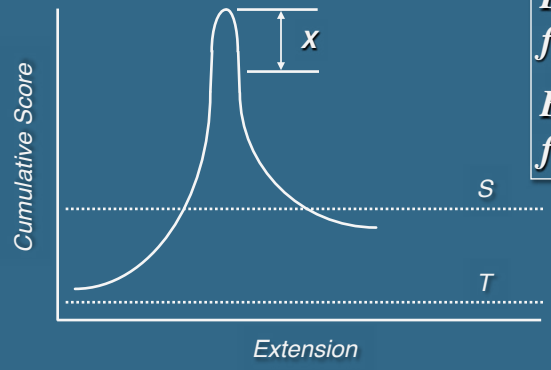

Extension
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L TP G R++ +W+ +P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
X
Significance decay
• mismatches
• gap penalties
Cumulative Score
S
T
Extension
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research


Scores and Probabilities
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L TP G R++ +W+ +P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
E = kmNe-λS
m # letters in query
N # letters in database
mN size of search space
λS normalized score
k minor constant
X
S
T
Cumulative Score
Extension
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research


Scores and Probabilities
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L TP G R++ +W+ +P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
E = kmNe-λS
Number of HSPs found purely by chance
Lower values signify higher similarity
X
S
T
Cumulative Score
Extension
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Scores and Probabilities

```
Query:   325  SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA  365
              +LA++L    TP G R++ +W+ +P+ D   + ER   + A
Sbjct:   290  TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA  330
```



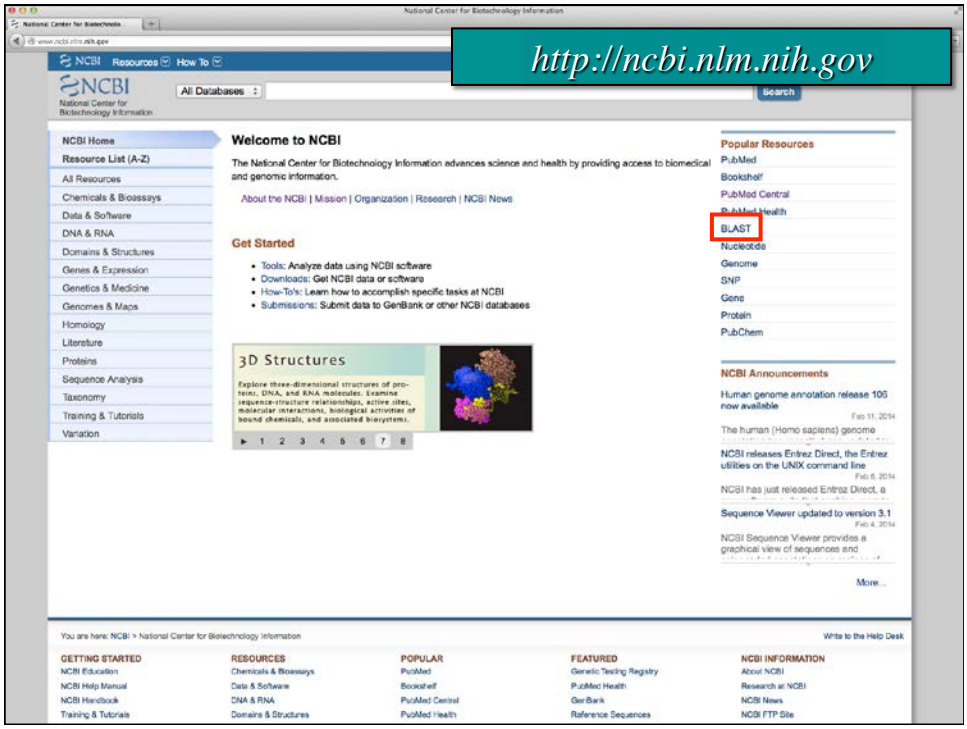
$E \leq 10^{-6}$ for nucleotides

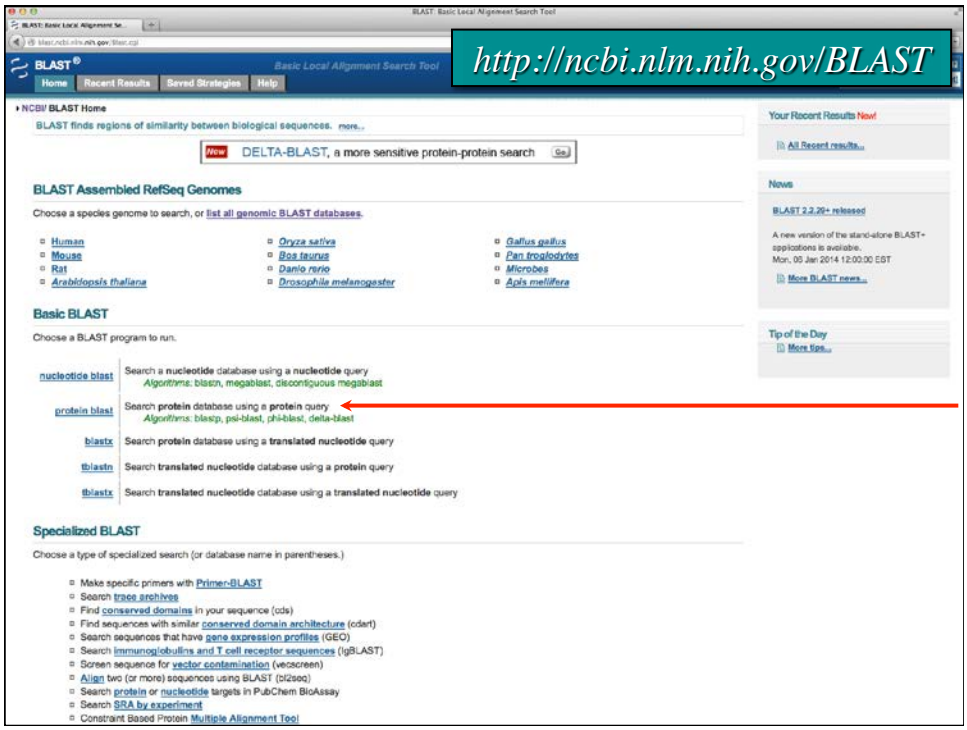
$E \leq 10^{-3}$ for proteins

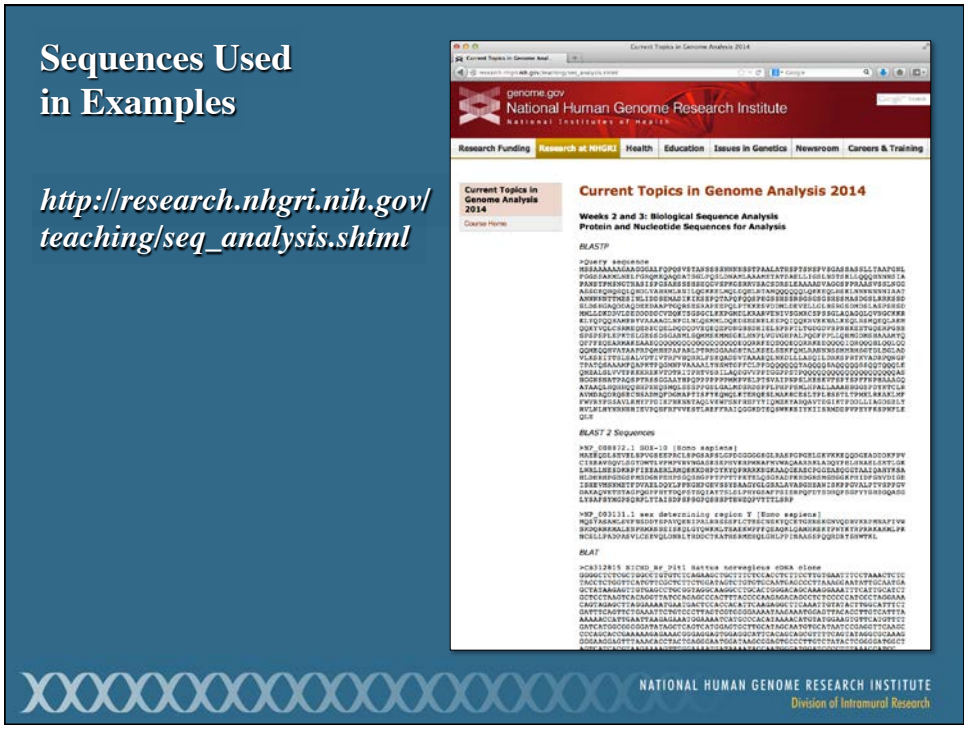
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

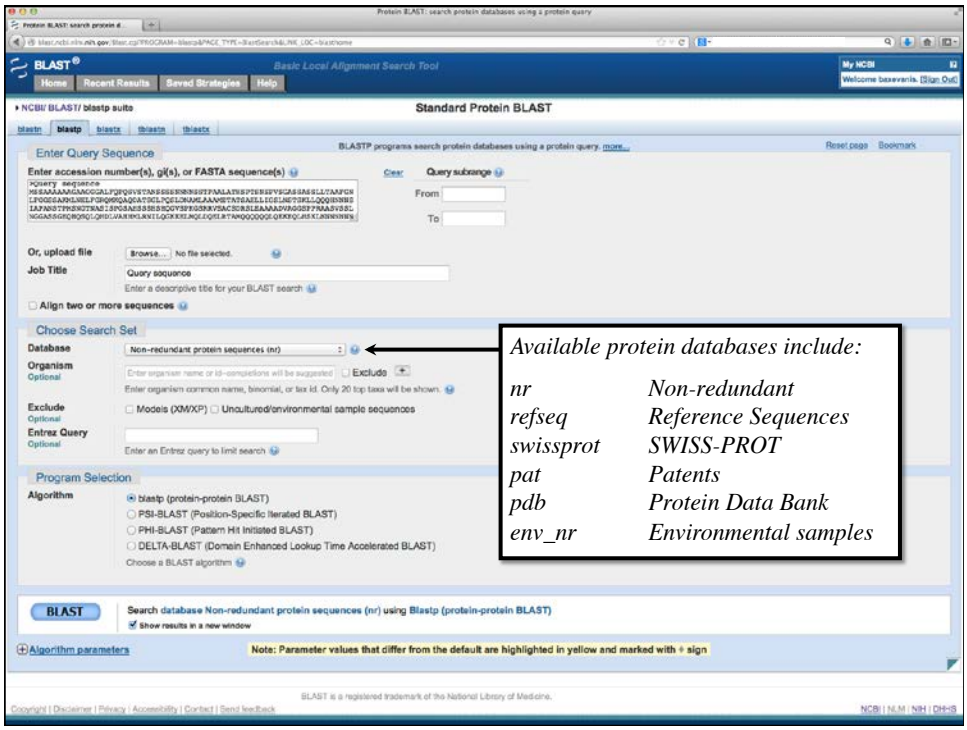
Using BLAST for Protein Similarity Searching

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research


http://ncbi.nlm.nih.gov
NCBI Resources How To
NCBI
National Center for Biotechnology Information
All Databases
Search
NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation
Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
About the NCBI | Mission | Organization | Research | NCBI News
Get Started
Tools: Analyze data using NCBI software
Downloads: Get NCBI data or software
How-To's: Learn how to accomplish specific tasks at NCBI
Submissions: Submit data to GenBank or other NCBI databases
3D Structures
Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem
NCBI Announcements
Human genome annotation release 106 now available
NCBI releases Entrez Direct, the Entrez utilities on the UNIX command line
Sequence Viewer updated to version 3.1
More...
GETTING STARTED
RESOURCES
POPULAR
FEATURED
NCBI INFORMATION


http://ncbi.nlm.nih.gov/BLAST
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI BLAST Home
BLAST finds regions of similarity between biological sequences. more...
New DELTA-BLAST, a more sensitive protein-protein search Go
BLAST Assembled RefSeq Genomes
Choose a species genome to search, or list all genomic BLAST databases.
Human
Mouse
Rat
Arabidopsis thaliana
Oryza sativa
Bos taurus
Danio rerio
Drosophila melanogaster
Gallus gallus
Pan troglodytes
Microbes
Apis mellifera
Basic BLAST
Choose a BLAST program to run.
nucleotide blast
Search a nucleotide database using a nucleotide query
Algorithms: blastn, megablast, discontiguous megablast
protein blast
Search protein database using a protein query
Algorithms: blastp, psi-blast, phi-blast, delta-blast
blastx
Search protein database using a translated nucleotide query
tblastn
Search translated nucleotide database using a protein query
tblastx
Search translated nucleotide database using a translated nucleotide query
Specialized BLAST
Choose a type of specialized search (or database name in parentheses.)
Make specific primers with Primer-BLAST
Search trace archives
Find conserved domains in your sequence (cds)
Find sequences with similar conserved domain architecture (cdart)
Search sequences that have gene expression profiles (GEO)
Search immunoglobulins and T cell receptor sequences (IgBLAST)
Screen sequence for vector contamination (vecscreen)
Align two (or more) sequences using BLAST (bl2seq)
Search protein or nucleotide targets in PubChem BioAssay
Search SRA by experiment
Constraint Based Protein Multiple Alignment Tool
Your Recent Results New!
All Recent results...
News
BLAST 2.2.29+ released
A new version of the stand-alone BLAST+ applications is available.
Mon, 06 Jan 2014 12:00:00 EST
More BLAST news...
Tip of the Day
More tips...


Sequences Used in Examples
http://research.nhgri.nih.gov/teaching/seq_analysis.shtml
genome.gov
National Human Genome Research Institute
Research Funding
Research at NHGRI
Health
Education
Issues in Genetics
Newsroom
Careers & Training
Current Topics in Genome Analysis 2014
Current Topics in Genome Analysis 2014
Weeks 2 and 3: Biological Sequence Analysis
Protein and Nucleotide Sequences for Analysis
BLASTP
BLAST 2 Sequences
BLAT
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research


BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
Standard Protein BLAST
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Query subrange
From
To
Or, upload file
Job Title
Query sequence
Align two or more sequences
Choose Search Set
Database
Non-redundant protein sequences (nr)
Organism
Optional
Exclude
Models (XM/XP)
Uncultured/environmental sample sequences
Entrez Query
Optional
Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with + sign
Available protein databases include:
nr Non-redundant
refseq Reference Sequences
swissprot SWISS-PROT
pat Patents
pdb Protein Data Bank
env_nr Environmental samples

NCBI RefSeq Database

- *Goal:* Provide a single reference sequence for each molecule of the central dogma (DNA, mRNA, and protein)
- Distinguishing features
  - Non-redundancy
  - Updates to reflect the current knowledge of sequence data and biology
  - Includes biological attributes of the gene, gene transcript, or protein
  - Encompasses a wide taxonomic range, with primary focus on mammalian and human species
  - Ongoing updates and curation (both automated and manual review), with review status indicated on each record

Pruitt et al., Nucleic Acids Res. 42: D756-D763, 2014

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

RefSeq Accession Number Prefixes

From curation of GenBank entries:

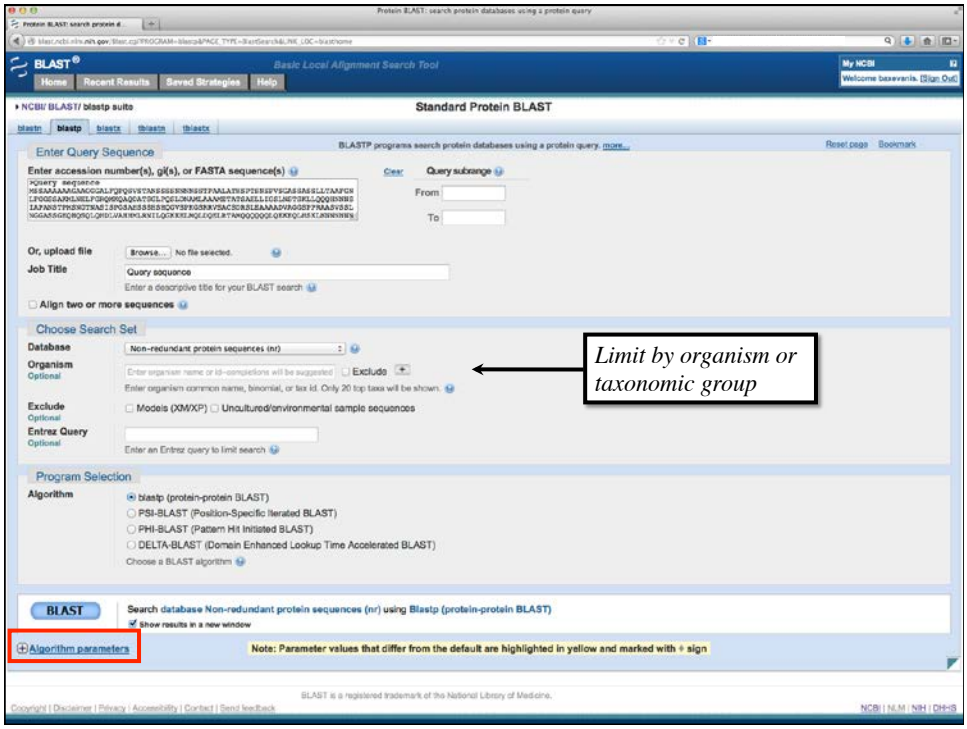
NT_	Genomic contigs
NM_	mRNAs
NP_	Proteins
NR_	Non-coding transcripts

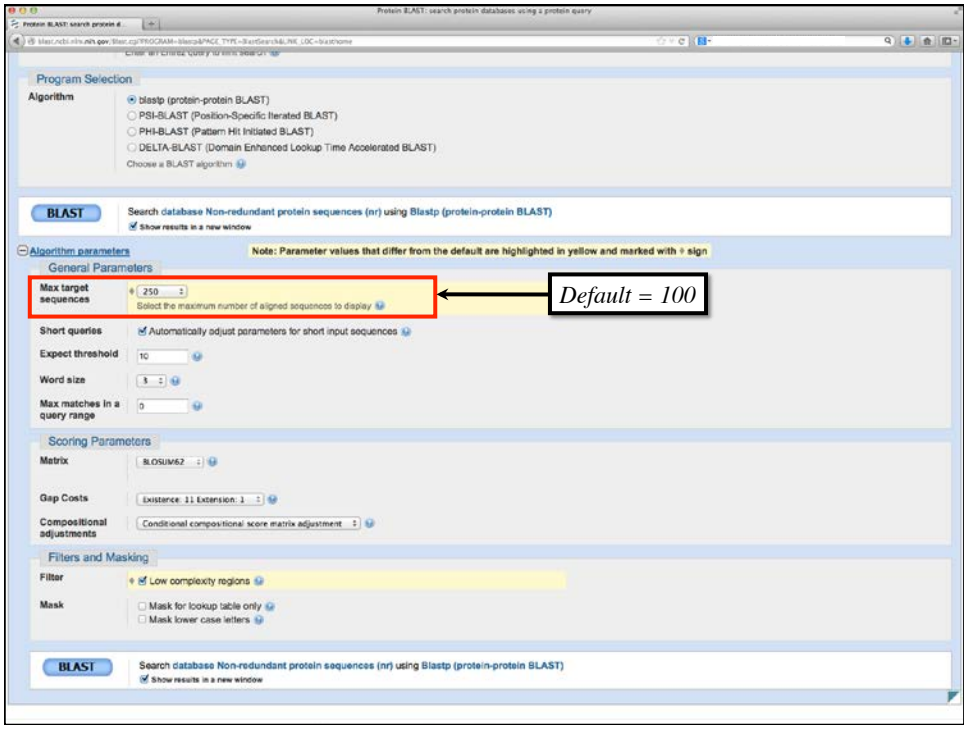
From genome annotation:

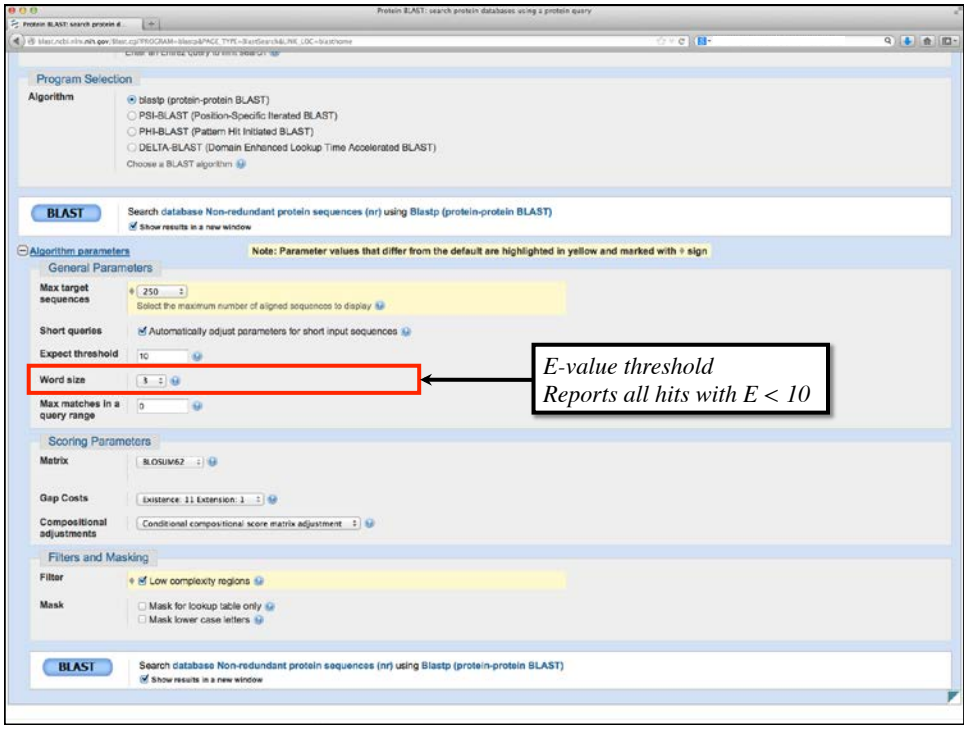
XM_	Model mRNA
XP_	Model proteins

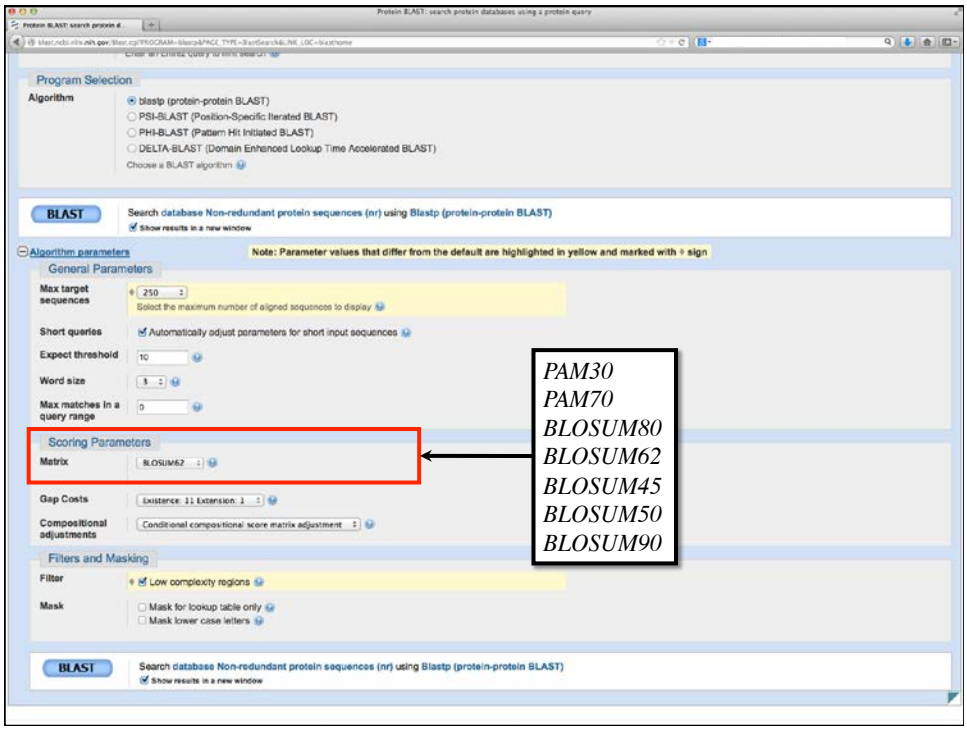
Complete list of molecule types in Chapter 18 of the NCBI Handbook
http://ncbi.nlm.nih.gov/books/NBK21091

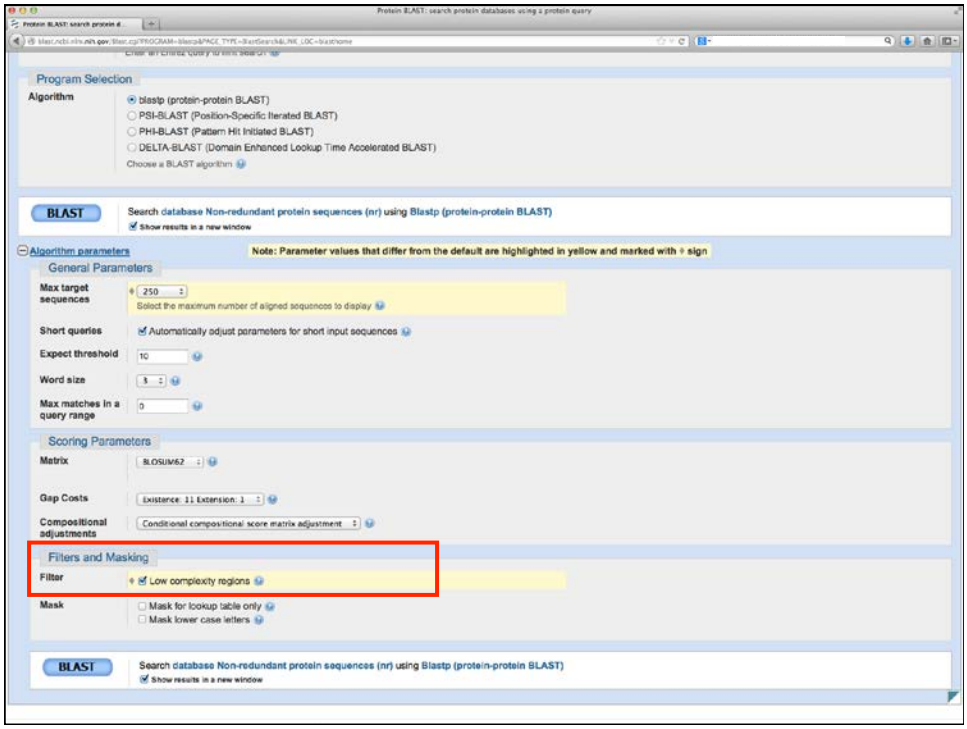
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research


Standard Protein BLAST
Enter Query Sequence
Choose Search Set
Program Selection
Limit by organism or taxonomic group
BLAST
Algorithm parameters


Program Selection
BLAST
Algorithm parameters
General Parameters
Max target sequences
Default = 100
Scoring Parameters
Filters and Masking

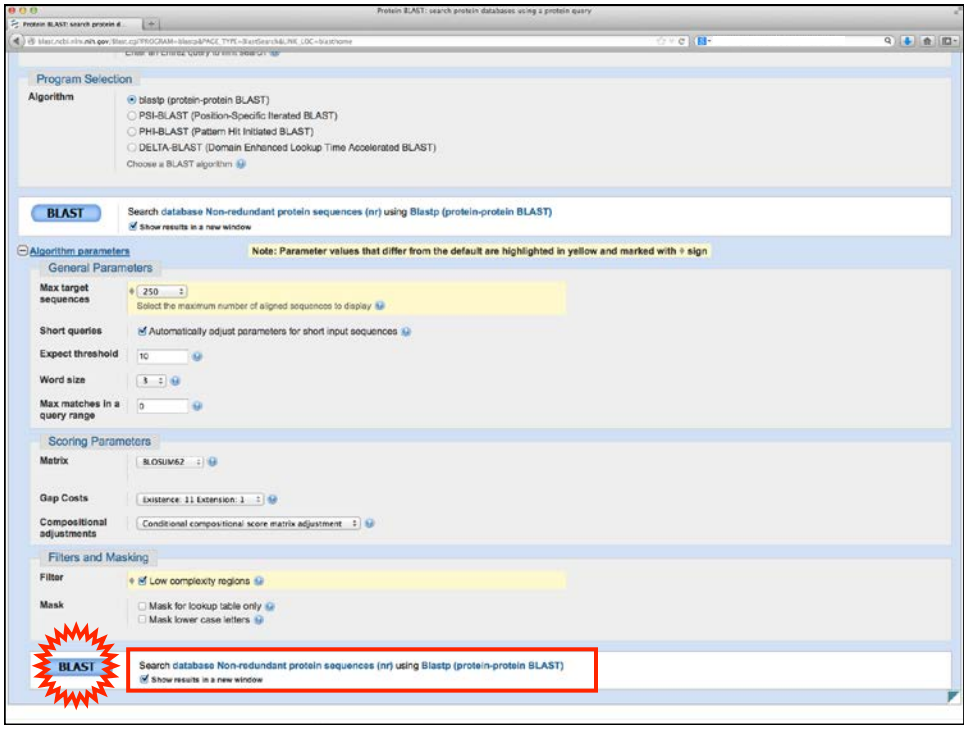

Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign
General Parameters
Max target sequences
Short queries
Expect threshold
Word size
Max matches in a query range
Scoring Parameters
Matrix
Gap Costs
Compositional adjustments
Filters and Masking
Filter
Mask
E-value threshold
Reports all hits with E < 10

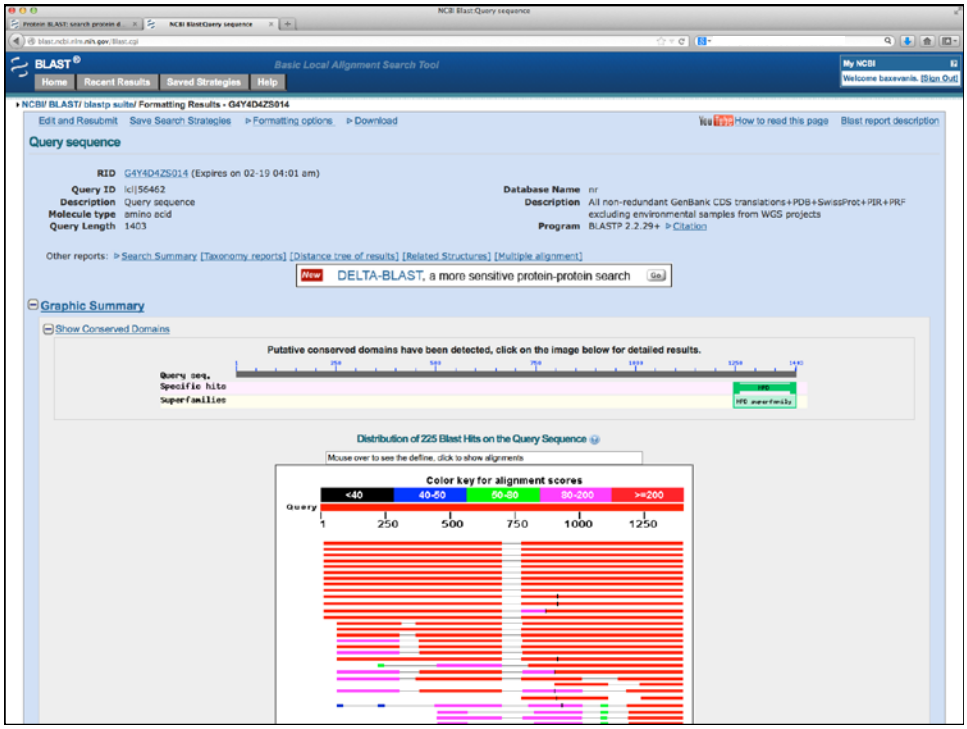

Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign
General Parameters
Max target sequences
Short queries
Expect threshold
Word size
Max matches in a query range
Scoring Parameters
Matrix
Gap Costs
Compositional adjustments
Filters and Masking
Filter
Mask
PAM30
PAM70
BLOSUM80
BLOSUM62
BLOSUM45
BLOSUM50
BLOSUM90

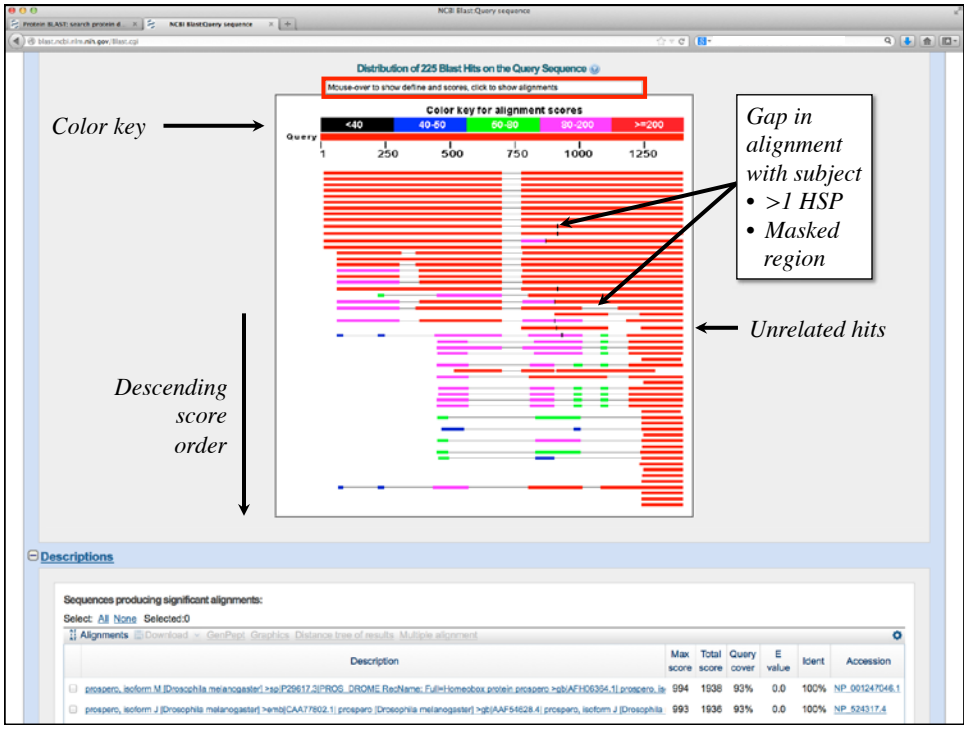


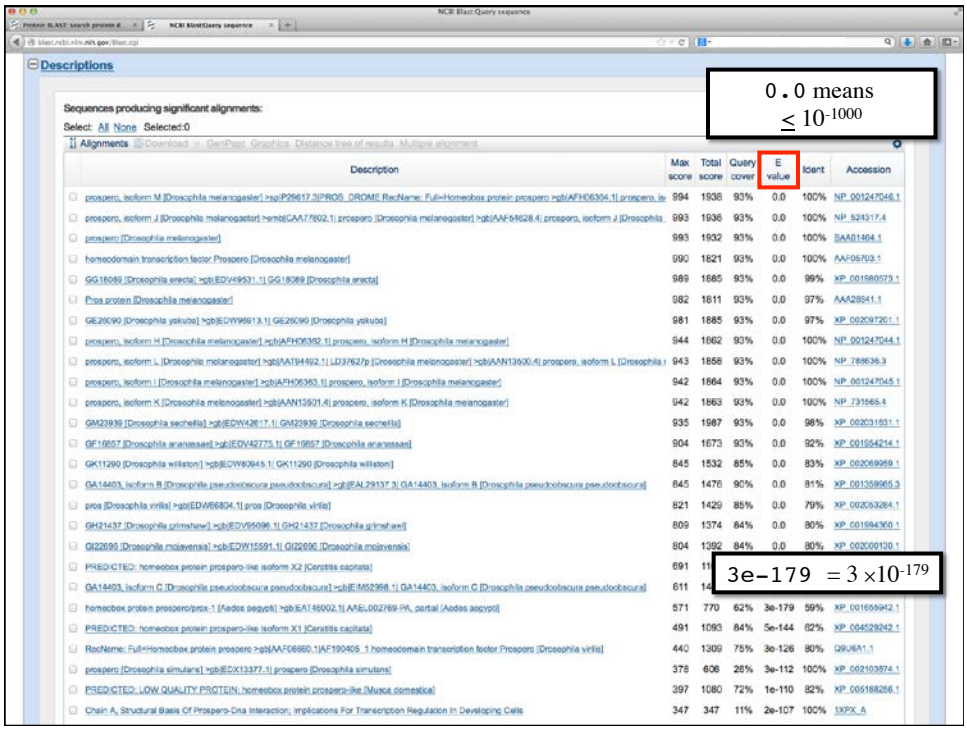
Low-Complexity Regions

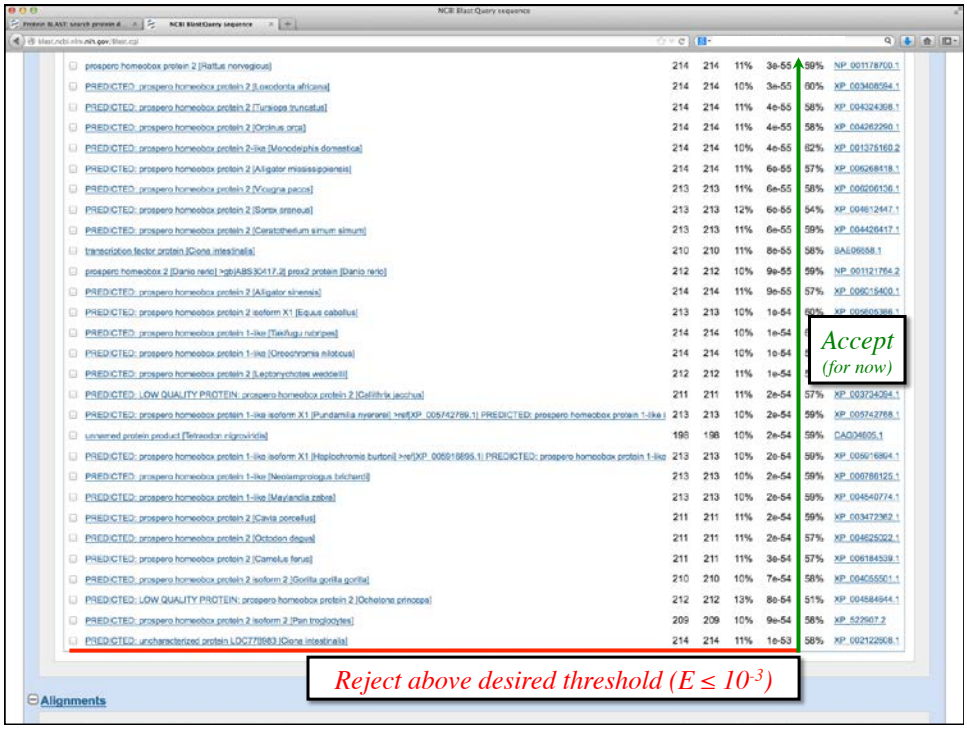
- Defined as regions of "biased composition"
  - Homopolymeric runs
  - Short-period repeats
  - Subtle over-representation of several residues
- May confound sequence analysis
  - BLAST relies on uniformly-distributed amino acid frequencies
  - Often lead to false positives
- Filtering is advised (but *not* enabled by default)

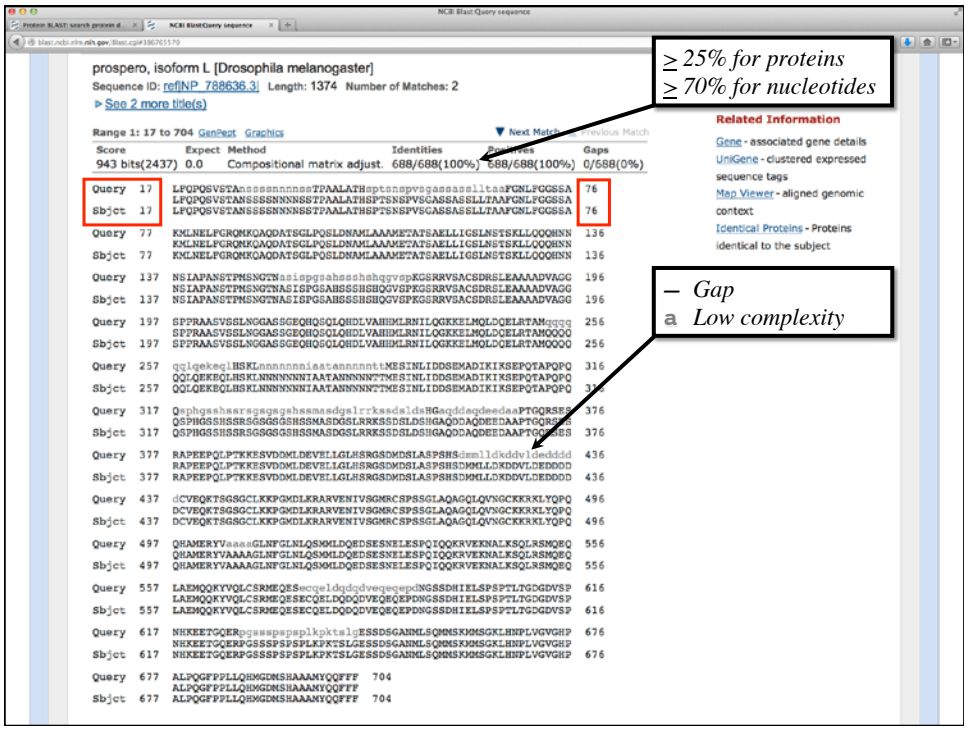

Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with + sign
General Parameters
Max target sequences
Short queries
Expect threshold
Word size
Max matches in a query range
Scoring Parameters
Matrix
Gap Costs
Compositional adjustments
Filters and Masking
Filter
Low complexity regions
Mask
Mask for lookup table only
Mask lower case letters
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

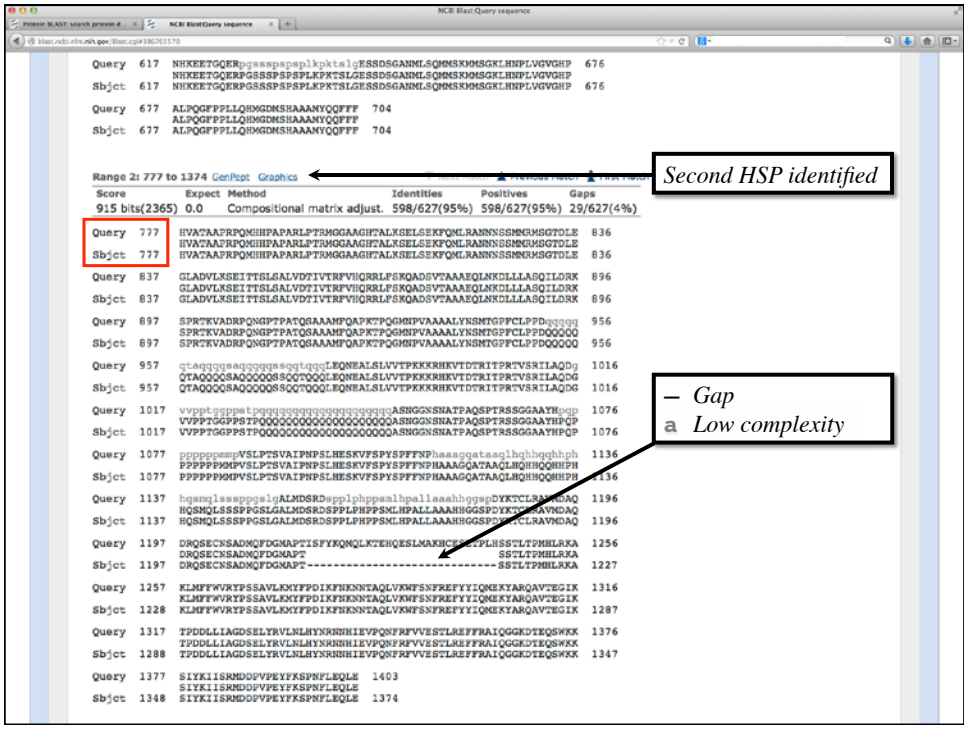

BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
Edit and Resubmit
Save Search Strategies
Formatting options
Download
How to read this page
Blast report description
Query sequence
RID
Query ID
Description
Molecule type
Query Length
Database Name
Program
Other reports:
DELTA-BLAST, a more sensitive protein-protein search
Graphic Summary
Show Conserved Domains
Putative conserved domains have been detected, click on the image below for detailed results.
Distribution of 225 Blast Hits on the Query Sequence
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
250
500
750
1000
1250

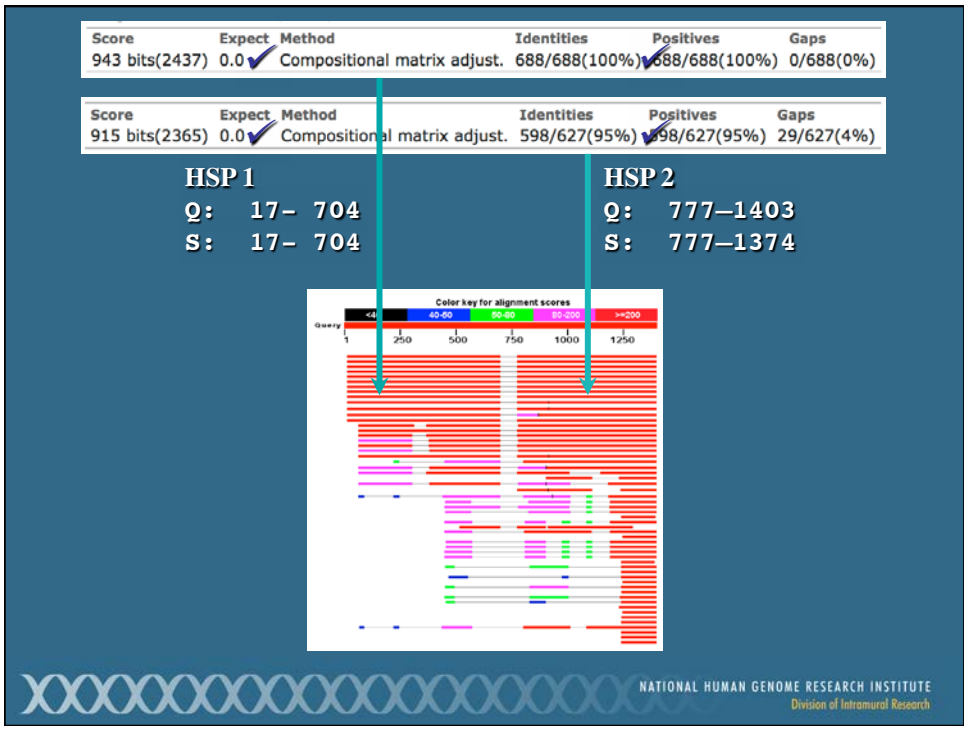

Distribution of 225 Blast Hits on the Query Sequence
Mouse-over to show define and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
250
500
750
1000
1250
Color key
Gap in alignment with subject
• >1 HSP
• Masked region
Unrelated hits
Descending score order
Descriptions
Sequences producing significant alignments:
Select: All None Selected:0
Alignments
Description
Max score
Total score
Query cover
E value
Ident
Accession


Descriptions
Sequences producing significant alignments:
Select: All None Selected:0
Alignments
0.0 means ≤ 10-1000
Description
Max score
Total score
Query cover
E value
Ident
Accession
3e-179 = 3 ×10-179


Accept
(for now)
Reject above desired threshold ($E \leq 10^{-3}$)
Alignments


≥ 25% for proteins
≥ 70% for nucleotides
– Gap
a Low complexity
Related Information
Gene - associated gene details
UniGene - clustered expressed sequence tags
Map Viewer - aligned genomic context
Identical Proteins - Proteins identical to the subject


Second HSP identified
– Gap
a Low complexity


Score Expect Method Identities Positives Gaps
943 bits(2437) 0.0 Compositional matrix adjust. 688/688(100%) 688/688(100%) 0/688(0%)
Score Expect Method Identities Positives Gaps
915 bits(2365) 0.0 Compositional matrix adjust. 598/627(95%) 598/627(95%) 29/627(4%)
HSP 1
Q: 17– 704
S: 17– 704
HSP 2
Q: 777–1403
S: 777–1374
Color key for alignment scores
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

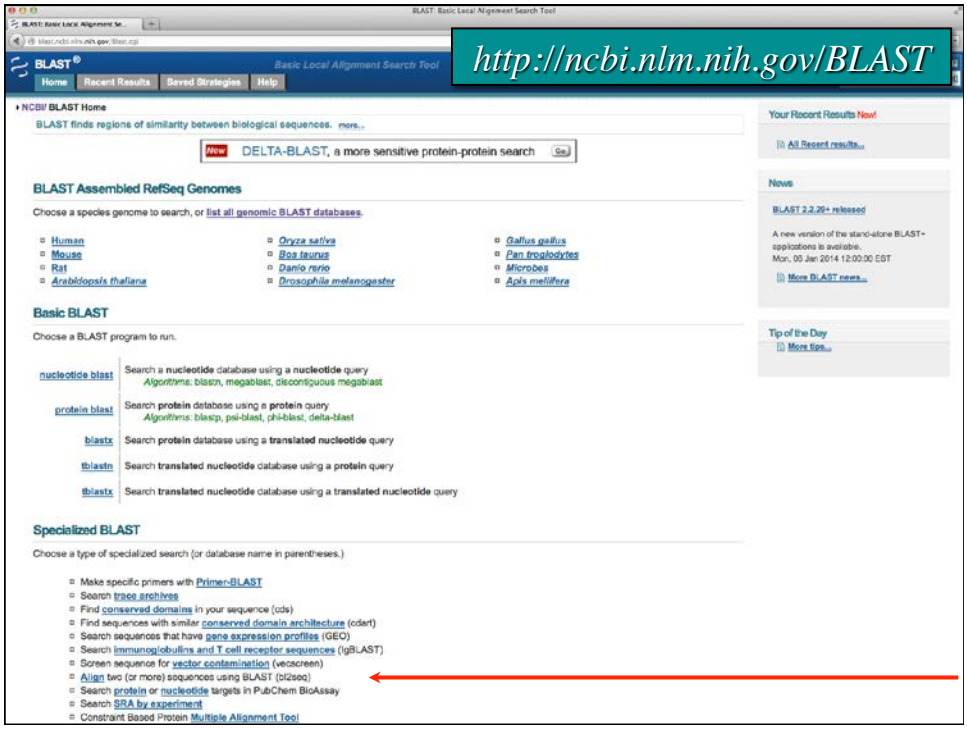
Suggested BLAST Cutoffs

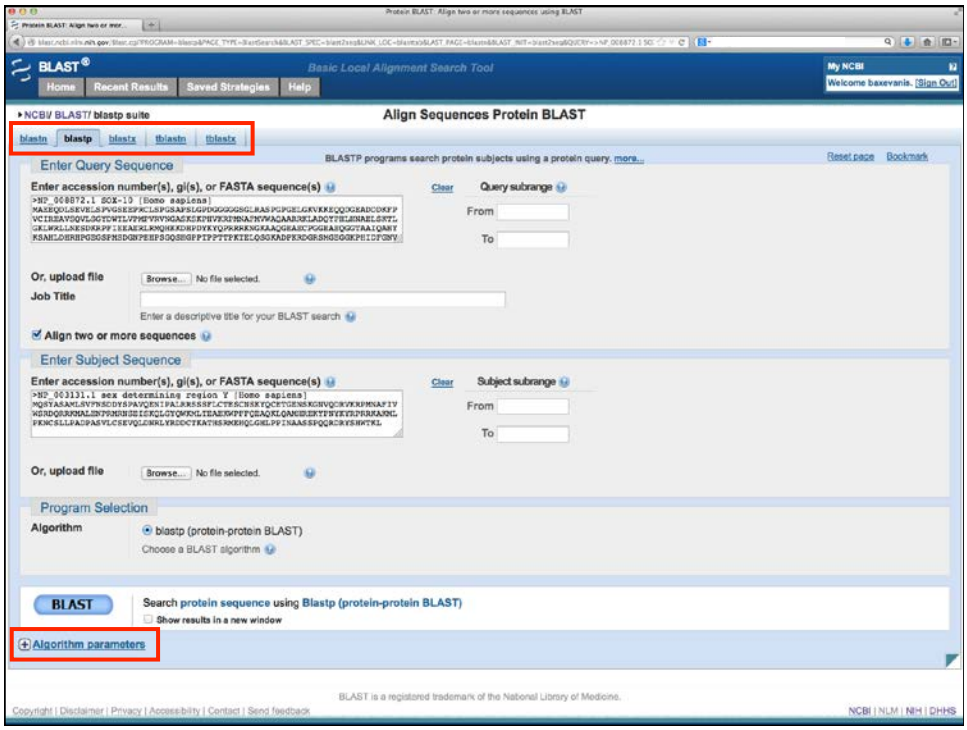
	E-value	Sequence Identity
Nucleotide	$\leq 10^{-6}$	$\geq 70\%$
Protein	$\leq 10^{-3}$	$\geq 25\%$

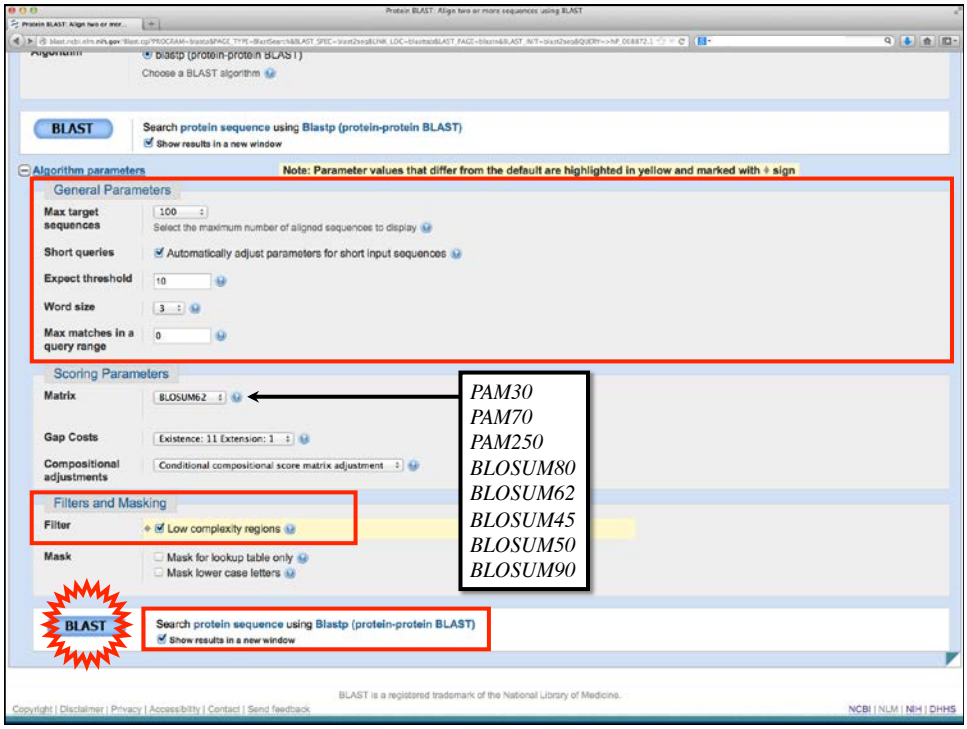
- ***Do not use these cutoffs blindly!***
- ***Pay attention to alignments on either side of the dividing line***
- ***Do not ignore biology!***

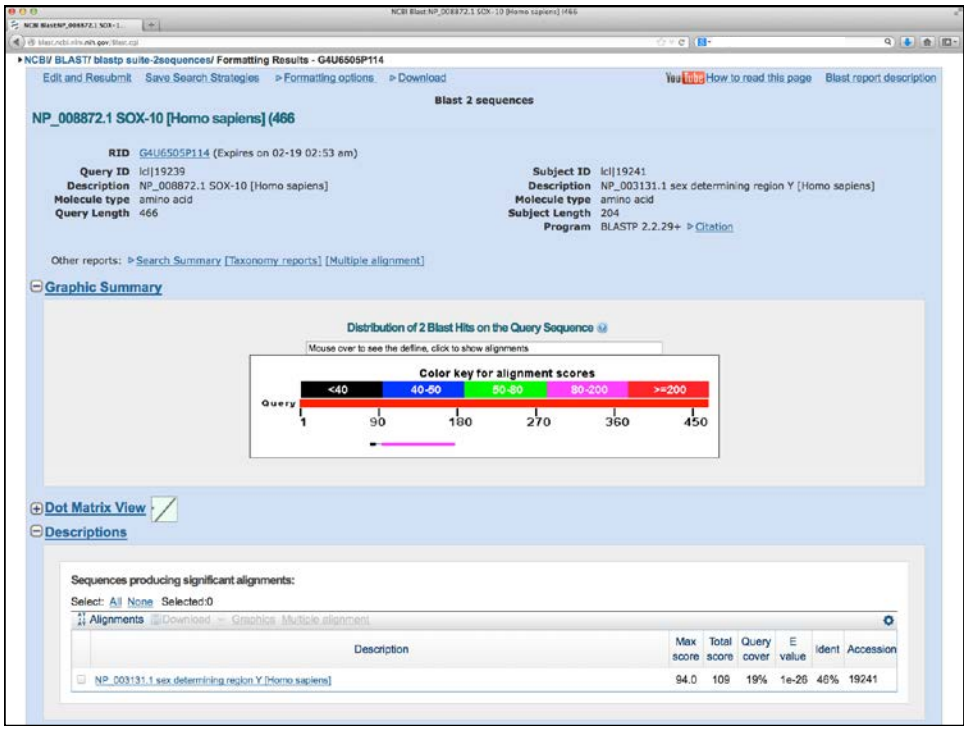
BLAST 2 Sequences

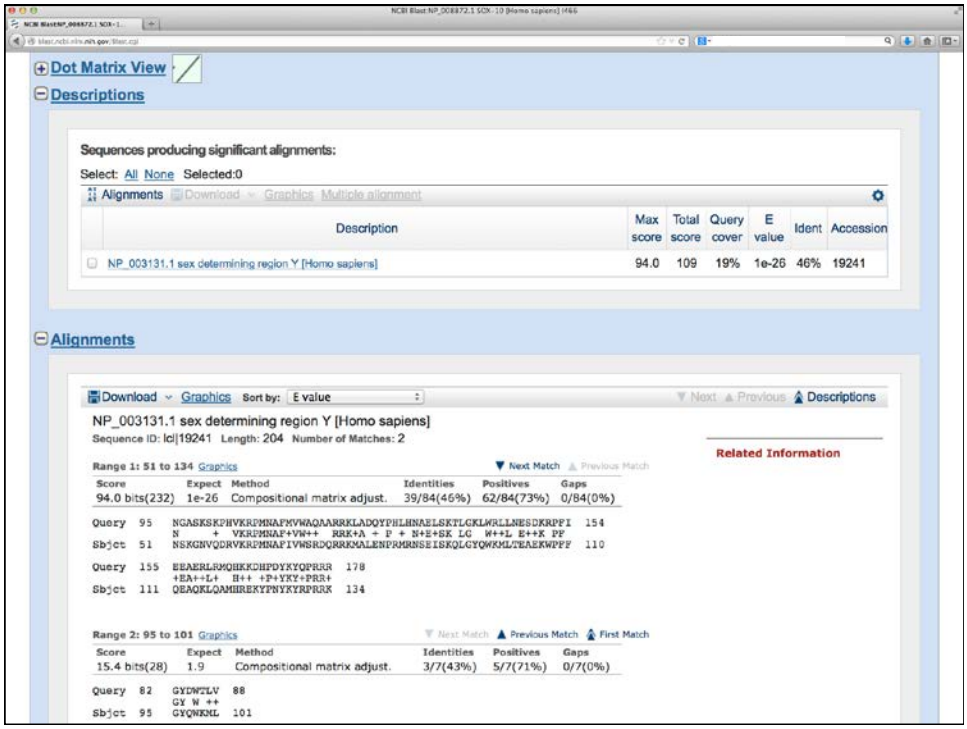
- Finds local alignments between two protein or nucleotide sequences of interest
- All BLAST programs available
- Select BLOSUM and PAM matrices available for protein comparisons
- Same affine gap costs (adjustable)
- Input sequences can be masked

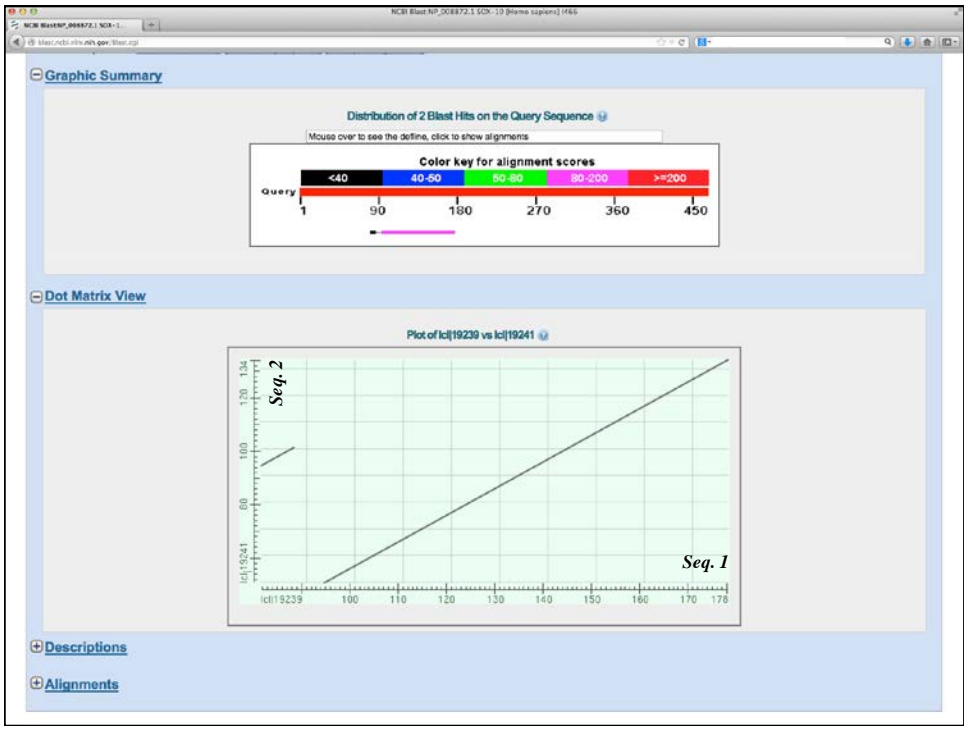

http://ncbi.nlm.nih.gov/BLAST
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI BLAST Home
BLAST finds regions of similarity between biological sequences. more...
New
DELTA-BLAST, a more sensitive protein-protein search
Go
BLAST Assembled RefSeq Genomes
Choose a species genome to search, or list all genomic BLAST databases.
Human
Mouse
Rat
Arabidopsis thaliana
Oryza sativa
Bos taurus
Danio rerio
Drosophila melanogaster
Gallus gallus
Pan troglodytes
Microbes
Apis mellifera
Basic BLAST
Choose a BLAST program to run.
nucleotide blast
Search a nucleotide database using a nucleotide query
Algorithms: blastn, megablast, discontiguous megablast
protein blast
Search protein database using a protein query
Algorithms: blastp, psi-blast, phi-blast, delta-blast
blastx
Search protein database using a translated nucleotide query
tblastn
Search translated nucleotide database using a protein query
tblastx
Search translated nucleotide database using a translated nucleotide query
Specialized BLAST
Choose a type of specialized search (or database name in parentheses.)
Make specific primers with Primer-BLAST
Search trace archives
Find conserved domains in your sequence (cds)
Find sequences with similar conserved domain architecture (cdart)
Search sequences that have gene expression profiles (GEO)
Search immunoglobulins and T cell receptor sequences (IgBLAST)
Screen sequence for vector contamination (vecscreen)
Align two (or more) sequences using BLAST (bl2seq)
Search protein or nucleotide targets in PubChem BioAssay
Search SRA by experiment
Constraint Based Protein Multiple Alignment Tool
Your Recent Results New!
All Recent results...
News
BLAST 2.2.29+ released
A new version of the stand-alone BLAST+ applications is available.
Mon, 06 Jan 2014 12:00:00 EST
More BLAST news...
Tip of the Day
More tips...


BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
Welcome baxevanis. [Sign Out]
NCBI BLAST/ blastp suite
Align Sequences Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein subjects using a protein query. more...
Reset page
Bookmark
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
Or, upload file
Browse...
No file selected.
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Enter Subject Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Subject subrange
From
To
Or, upload file
Browse...
No file selected.
Program Selection
Algorithm
blastp (protein-protein BLAST)
Choose a BLAST algorithm
BLAST
Search protein sequence using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters
BLAST is a registered trademark of the National Library of Medicine.
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback
NCBI | NLM | NIH | DHHS


Choose a BLAST algorithm
BLAST
Search protein sequence using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with + sign
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
3
Max matches in a query range
0
Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Compositional adjustments
Conditional compositional score matrix adjustment
Filters and Masking
Filter
Low complexity regions
Mask
Mask for lookup table only
Mask lower case letters
PAM30
PAM70
PAM250
BLOSUM80
BLOSUM62
BLOSUM45
BLOSUM50
BLOSUM90
BLAST
Search protein sequence using Blastp (protein-protein BLAST)
Show results in a new window
BLAST is a registered trademark of the National Library of Medicine.
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback
NCBI | NLM | NIH | DHHS

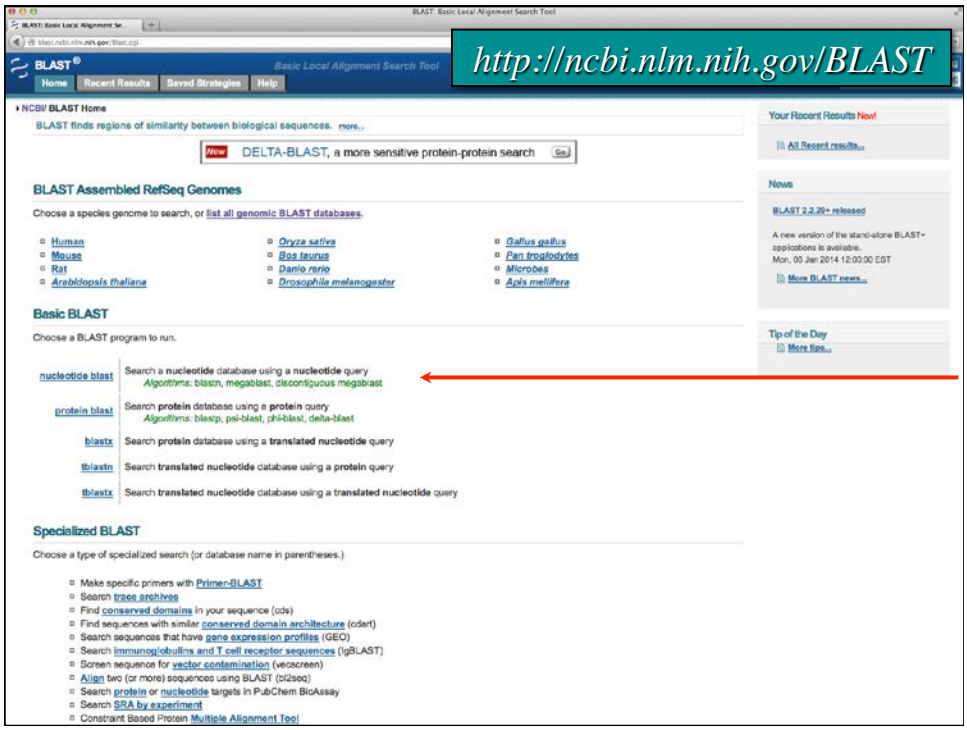

NCBI/ BLAST/ blastp suite-2sequences/ Formatting Results - G4U6505P114
Edit and Resubmit
Save Search Strategies
Formatting options
Download
How to read this page
Blast report description
Blast 2 sequences
NP_008872.1 SOX-10 [Homo sapiens] (466
RID G4U6505P114 (Expires on 02-19 02:53 am)
Query ID lcl|19239
Description NP_008872.1 SOX-10 [Homo sapiens]
Molecule type amino acid
Query Length 466
Subject ID lcl|19241
Description NP_003131.1 sex determining region Y [Homo sapiens]
Molecule type amino acid
Subject Length 204
Program BLASTP 2.2.29+ Citation
Other reports: Search Summary [Taxonomy reports] [Multiple alignment]
Graphic Summary
Distribution of 2 Blast Hits on the Query Sequence
Mouse over to see the define, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
90
180
270
360
450
Dot Matrix View
Descriptions
Sequences producing significant alignments:
Select: All None Selected:0
Alignments Download Graphics Multiple alignment
Description
Max score
Total score
Query cover
E value
Ident
Accession
NP_003131.1 sex determining region Y [Homo sapiens]
94.0
109
19%
1e-26
46%
19241

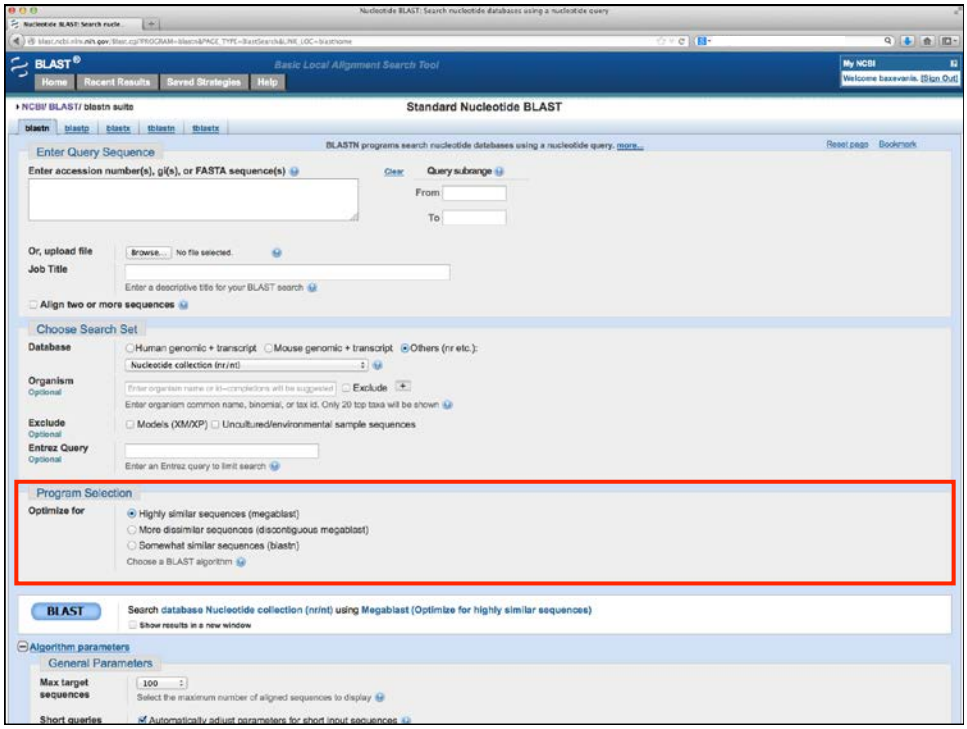

Dot Matrix View
Descriptions
Sequences producing significant alignments:
Select: All None Selected:0
Alignments Download Graphics Multiple alignment
Description Max score Total score Query cover E value Ident Accession
NP_003131.1 sex determining region Y [Homo sapiens] 94.0 109 19% 1e-26 46% 19241
Alignments
Download Graphics Sort by: E value
Next Previous Descriptions
NP_003131.1 sex determining region Y [Homo sapiens]
Sequence ID: lcl|19241 Length: 204 Number of Matches: 2
Related Information
Range 1: 51 to 134 Graphics
Next Match Previous Match
Score Expect Method Identities Positives Gaps
94.0 bits(232) 1e-26 Compositional matrix adjust. 39/84(46%) 62/84(73%) 0/84(0%)
Query 95 NGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTLGKLWRLLNESDKRPFI 154
N + VKRPMNAF+VW++ RRK+A + P + N+E+SK LG W++L E++K PF
Sbjct 51 NSKGNVQDRVKRPMNAFIVWSRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFF 110
Query 155 EEAERLRMQHKKDHPDYKYQPRRR 178
+EA++L+ H++ +P+YKY+PRR+
Sbjct 111 QEAQKLQAMHREKYPNYKYRPRRK 134
Range 2: 95 to 101 Graphics
Next Match Previous Match First Match
Score Expect Method Identities Positives Gaps
15.4 bits(28) 1.9 Compositional matrix adjust. 3/7(43%) 5/7(71%) 0/7(0%)
Query 82 GYDWTLV 88
GY W ++
Sbjct 95 GYQWKML 101


Graphic Summary
Distribution of 2 Blast Hits on the Query Sequence
Mouse over to see the define, click to show alignments
Color key for alignment scores
<40 40-50 50-80 80-200 >=200
Query
1 90 180 270 360 450
Dot Matrix View
Plot of lcl|19239 vs lcl|19241
Seq. 2
Seq. 1
lcl|19241
lcl|19239
Descriptions
Alignments

Nucleotide Similarity Searching MegaBLAST, BLASTN, and BLAT

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research





Nucleotide-Based BLAST Algorithms

	W	*+/–*	*Gaps*
Optimized for aligning very long and/or highly similar sequences (> 95%)			
MegaBLAST *(default)*	28	1, –2	Linear
Better for diverged sequences and/or cross-species comparisons (< 80%)			
Discontiguous MegaBLAST	11	2, –3	Affine
BLASTN	11	2, –3	Affine
Finding short, nearly exact matches (< 20 bases)			
BLASTN	7	2, –3	Affine

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

BLAT

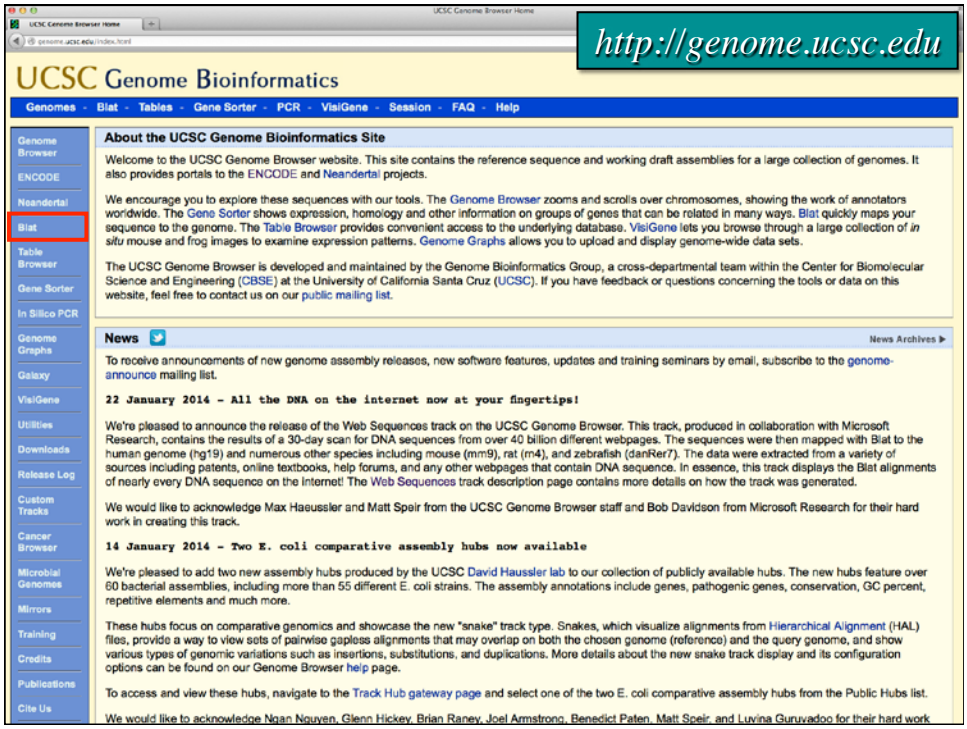
- "BLAST-Like Alignment Tool"
- Designed to rapidly align longer nucleotide sequences ($L \geq 40$) having $\geq 95\%$ sequence similarity
- Can find exact matches reliably down to $L = 33$
- Method of choice when looking for exact matches in nucleotide databases
- 500 times faster than BLAST for mRNA/DNA searches
- May miss divergent or shorter sequence alignments
- Can be used on protein sequences, but BLASTP is more efficient

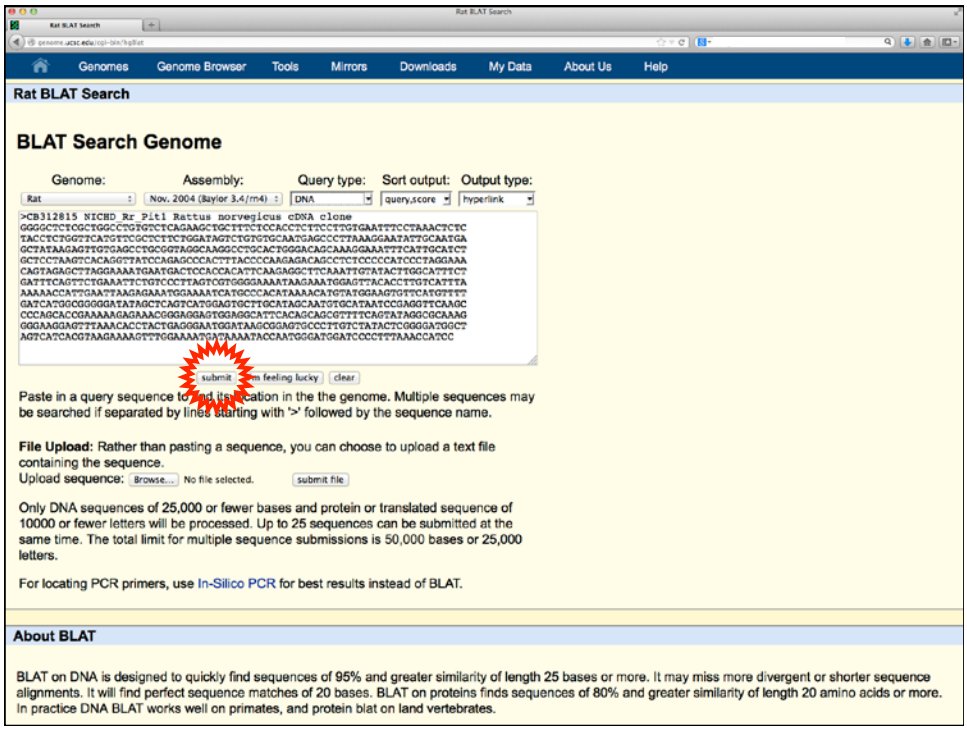
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

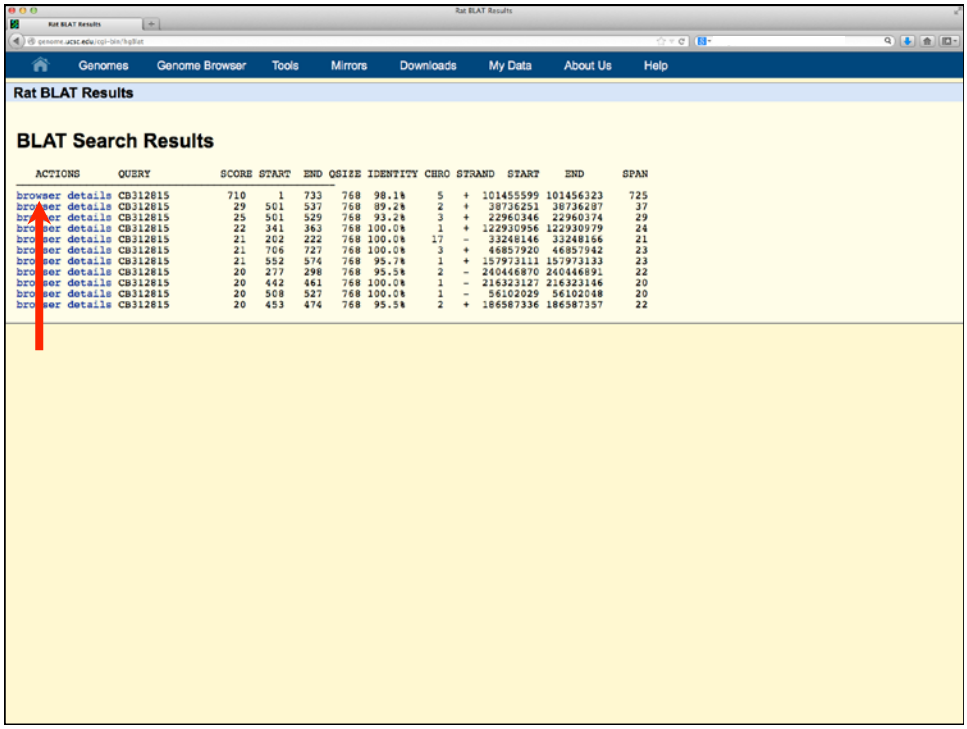
When to Use BLAT

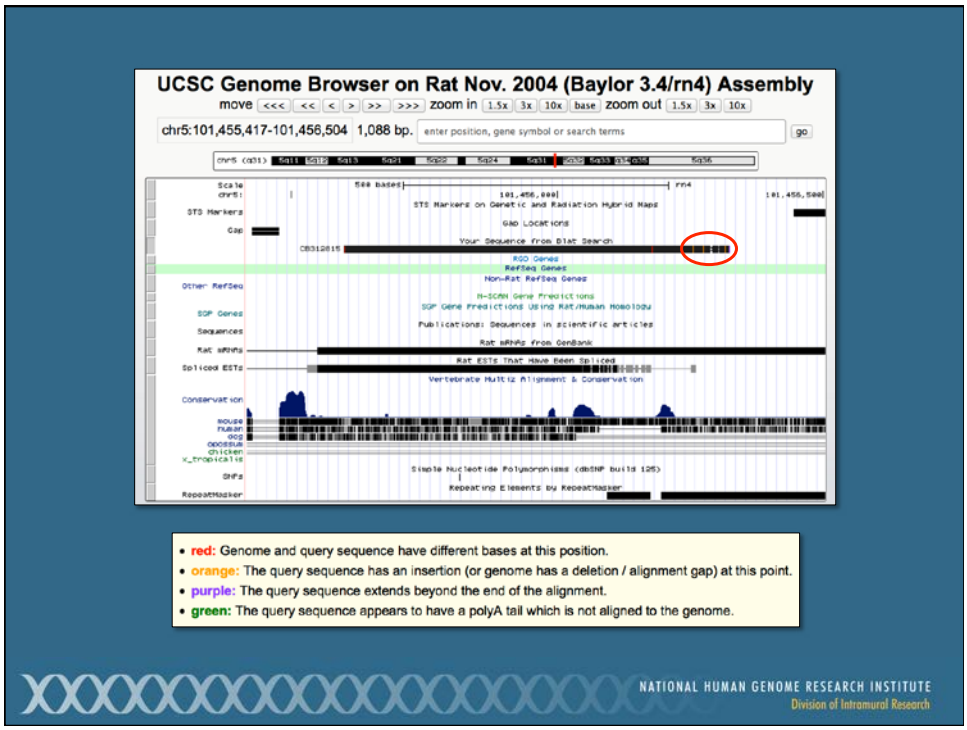
- To characterize an unknown gene or sequence fragment
  - Find its genomic coordinates
  - Determine gene structure (the presence and position of exons)
  - Identify markers of interest in the vicinity of a sequence
- To find highly similar (or identical) sequences
  - Alignment of mRNA sequences onto a genome assembly
  - Identification of gene family members
  - Cross-species alignment to identify putative homologs
- To display a specific sequence as a separate track within the UCSC Genome Browser

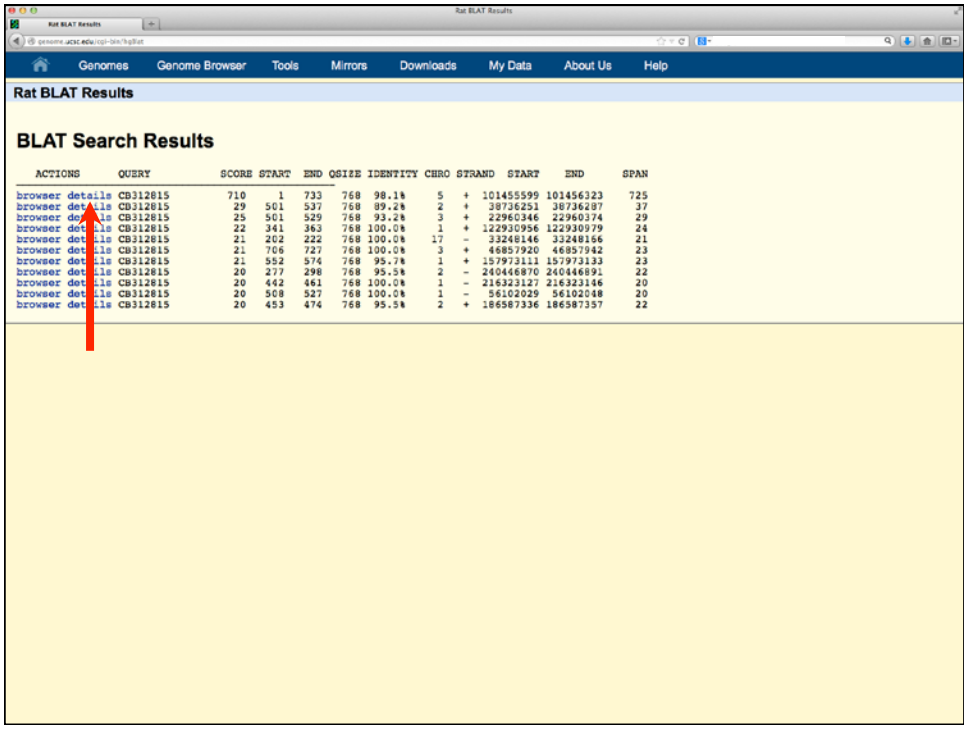
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

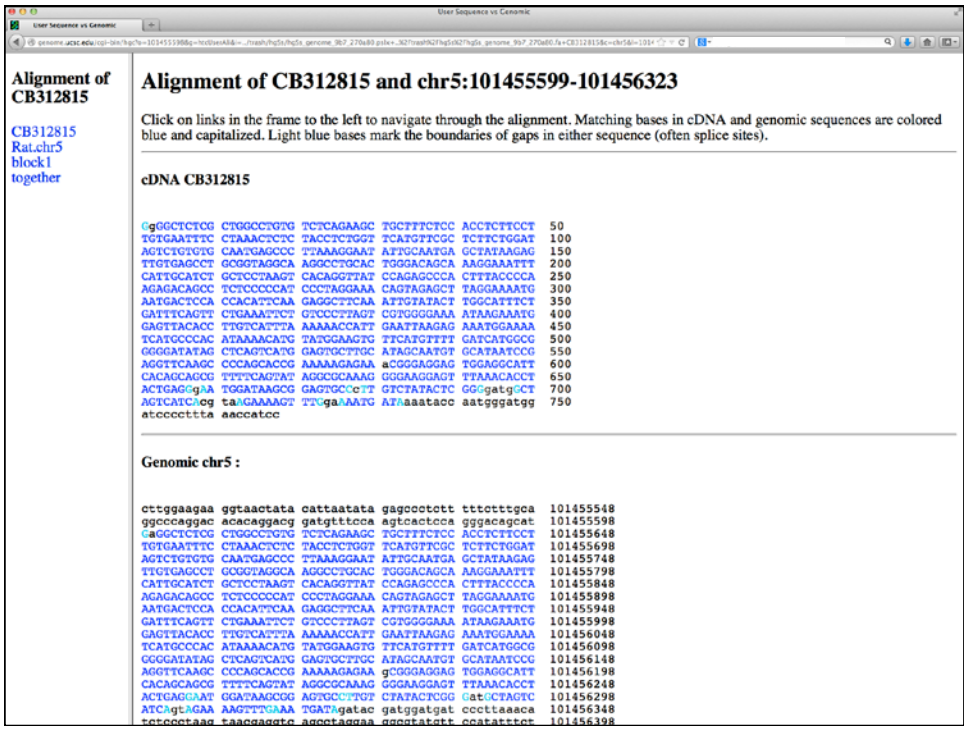

http://genome.ucsc.edu
UCSC Genome Bioinformatics
Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Session - FAQ - Help
Genome Browser
ENCODE
Neandertal
Blat
Table Browser
Gene Sorter
In Silico PCR
Genome Graphs
Galaxy
VisiGene
Utilities
Downloads
Release Log
Custom Tracks
Cancer Browser
Microbial Genomes
Mirrors
Training
Credits
Publications
Cite Us
About the UCSC Genome Bioinformatics Site
News
News Archives

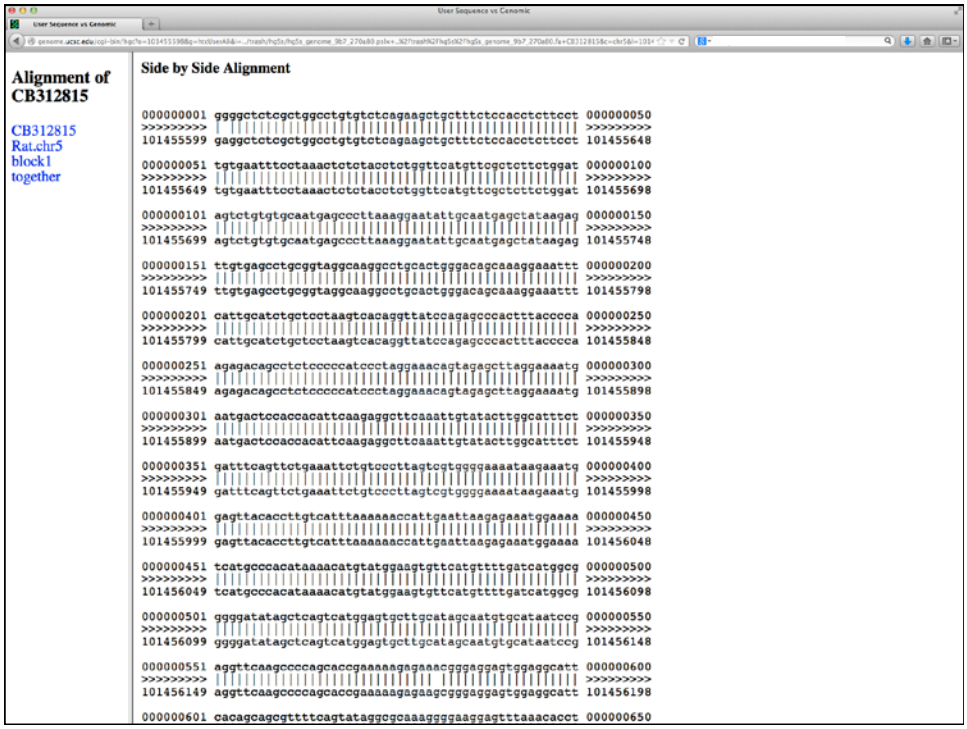

Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
About Us
Help
Rat BLAT Search
BLAT Search Genome
Genome:
Assembly:
Query type:
Sort output:
Output type:
submit
clear
About BLAT


Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
About Us
Help
Rat BLAT Results
BLAT Search Results
ACTIONS QUERY SCORE START END QSIZE IDENTITY CHRO STRAND START END SPAN
browser details CB312815 710 1 733 768 98.1% 5 + 101455599 101456323 725
browser details CB312815 29 501 537 768 89.2% 2 + 38736251 38736287 37
browser details CB312815 25 501 529 768 93.2% 3 + 22960346 22960374 29
browser details CB312815 22 341 363 768 100.0% 1 + 122930956 122930979 24
browser details CB312815 21 202 222 768 100.0% 17 - 33248146 33248166 21
browser details CB312815 21 706 727 768 100.0% 3 + 46857920 46857942 23
browser details CB312815 21 552 574 768 95.7% 1 + 157973111 157973133 23
browser details CB312815 20 277 298 768 95.5% 2 - 240446870 240446891 22
browser details CB312815 20 442 461 768 100.0% 1 - 216323127 216323146 20
browser details CB312815 20 508 527 768 100.0% 1 - 56102029 56102048 20
browser details CB312815 20 453 474 768 95.5% 2 + 186587336 186587357 22


UCSC Genome Browser on Rat Nov. 2004 (Baylor 3.4/rn4) Assembly
move zoom in zoom out
chr5:101,455,417-101,456,504 1,088 bp.
enter position, gene symbol or search terms
go
STS Markers
Gap
Your Sequence from Blat Search
RGD Genes
RefSeq Genes
Non-Rat RefSeq Genes
Other RefSeq
SGP Genes
Sequences
Rat mRNAs
Spliced ESTs
Conservation
Vertebrate Multiz Alignment & Conservation
SNPs
RepeatMasker
red: Genome and query sequence have different bases at this position.
orange: The query sequence has an insertion (or genome has a deletion / alignment gap) at this point.
purple: The query sequence extends beyond the end of the alignment.
green: The query sequence appears to have a polyA tail which is not aligned to the genome.
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research


Rat BLAT Results
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
About Us
Help
Rat BLAT Results
BLAT Search Results
ACTIONS QUERY SCORE START END QSIZE IDENTITY CHRO STRAND START END SPAN
browser details CB312815 710 1 733 768 98.1% 5 + 101455599 101456323 725
browser details CB312815 29 501 537 768 89.2% 2 + 38736251 38736287 37
browser details CB312815 25 501 529 768 93.2% 3 + 22960346 22960374 29
browser details CB312815 22 341 363 768 100.0% 1 + 122930956 122930979 24
browser details CB312815 21 202 222 768 100.0% 17 - 33248146 33248166 21
browser details CB312815 21 706 727 768 100.0% 3 + 46857920 46857942 23
browser details CB312815 21 552 574 768 95.7% 1 + 157973111 157973133 23
browser details CB312815 20 277 298 768 95.5% 2 - 240446870 240446891 22
browser details CB312815 20 442 461 768 100.0% 1 - 216323127 216323146 20
browser details CB312815 20 508 527 768 100.0% 1 - 56102029 56102048 20
browser details CB312815 20 453 474 768 95.5% 2 + 186587336 186587357 22


User Sequence vs Genomic
Alignment of CB312815
CB312815
Rat.chr5
block1
together
Alignment of CB312815 and chr5:101455599-101456323
Click on links in the frame to the left to navigate through the alignment. Matching bases in cDNA and genomic sequences are colored blue and capitalized. Light blue bases mark the boundaries of gaps in either sequence (often splice sites).
cDNA CB312815
GgGGCTCTCG CTGGCCTGTG TCTCAGAAGC TGCTTTCTCC ACCTCTTCCT 50
TGTGAATTTC CTAAACTCTC TACCTCTGGT TCATGTTCGC TCTTCTGGAT 100
AGTCTGTGTG CAATGAGCCC TTAAAGGAAT ATTGCAATGA GCTATAAGAG 150
TTGTGAGCCT GCGGTAGGCA AGGCCTGCAC TGGGACAGCA AAGGAAATTT 200
CATTGCATCT GCTCCTAAGT CACAGGTTAT CCAGAGCCCA CTTTACCCCA 250
AGAGACAGCC TCTCCCCCAT CCCTAGGAAA CAGTAGAGCT TAGGAAAATG 300
AATGACTCCA CCACATTCAA GAGGCTTCAA ATTGTATACT TGGCATTTCT 350
GATTTCAGTT CTGAAATTCT GTCCCTTAGT CGTGGGGAAA ATAAGAAATG 400
GAGTTACACC TTGTCATTTA AAAAACCATT GAATTAAGAG AAATGGAAAA 450
TCATGCCCAC ATAAAACATG TATGGAAGTG TTCATGTTTT GATCATGGCG 500
GGGGATATAG CTCAGTCATG GAGTGCTTGC ATAGCAATGT GCATAATCCG 550
AGGTTCAAGC CCCAGCACCG AAAAAGAGAA aCGGGAGGAG TGGAGGCATT 600
CACAGCAGCG TTTTCAGTAT AGGCGCAAAG GGGAAGGAGT TTAAACACCT 650
ACTGAGCgAA TGGATAAGCG GAGTGCCeTT GTCTATACTC GGCgatgCCT 700
AGTCATCAcg taAGAAAAGT TTGgaAAATG ATAaaatacc aatgggatgg 750
atccccttta aaccatcc
Genomic chr5 :
cttggaagaa ggtaactata cattaatata gagccctctt tttctttgca 101455548
ggcccaggac acacaggacg gatgtttcca agtcactcca gggacagcat 101455598
GaGGCTCTCG CTGGCCTGTG TCTCAGAAGC TGCTTTCTCC ACCTCTTCCT 101455648
TGTGAATTTC CTAAACTCTC TACCTCTGGT TCATGTTCGC TCTTCTGGAT 101455698
AGTCTGTGTG CAATGAGCCC TTAAAGGAAT ATTGCAATGA GCTATAAGAG 101455748
TTGTGAGCCT GCGGTAGGCA AGGCCTGCAC TGGGACAGCA AAGGAAATTT 101455798
CATTGCATCT GCTCCTAAGT CACAGGTTAT CCAGAGCCCA CTTTACCCCA 101455848
AGAGACAGCC TCTCCCCCAT CCCTAGGAAA CAGTAGAGCT TAGGAAAATG 101455898
AATGACTCCA CCACATTCAA GAGGCTTCAA ATTGTATACT TGGCATTTCT 101455948
GATTTCAGTT CTGAAATTCT GTCCCTTAGT CGTGGGGAAA ATAAGAAATG 101455998
GAGTTACACC TTGTCATTTA AAAAACCATT GAATTAAGAG AAATGGAAAA 101456048
TCATGCCCAC ATAAAACATG TATGGAAGTG TTCATGTTTT GATCATGGCG 101456098
GGGGATATAG CTCAGTCATG GAGTGCTTGC ATAGCAATGT GCATAATCCG 101456148
AGGTTCAAGC CCCAGCACCG AAAAAGAGAA gCGGGAGGAG TGGAGGCATT 101456198
CACAGCAGCG TTTTCAGTAT AGGCGCAAAG GGGAAGGAGT TTAAACACCT 101456248
ACTGAGGAAT GGATAAGCGG AGTGCCTTGT CTATACTCGG GatgCCTAGTC 101456298
ATCAgtAGAA AAGTTTGAAA TGATAgatac gatggatgat ccctaaaaca 101456348
tctcgctaag taacgaggtg aggctaggaa ggcgtatgtt ccatatttct 101456398



FASTA

- Identifies regions of local alignment
- Employs an approximation of the Smith-Waterman algorithm to determine the best alignment between two sequences
- Method is significantly different from that used by BLAST
- Online implementations at:

  http://fasta.bioch.virginia.edu
  http://www.ebi.ac.uk/fasta33


NIH
Intramural Research Program
Our Research Changes Lives
one program
many people
infinite possibilities
irp.nih.gov