

A Beginner's Guide to BLAST



Ann Hess


Andre Ptitsyn

Center for Bioinformatics

Colorado State University

1

Outline

- BLAST Basics
 - Example 1: Cross-Species Hybridization
 - Example 2: Leptin
 - Beyond the Basics
- 

2

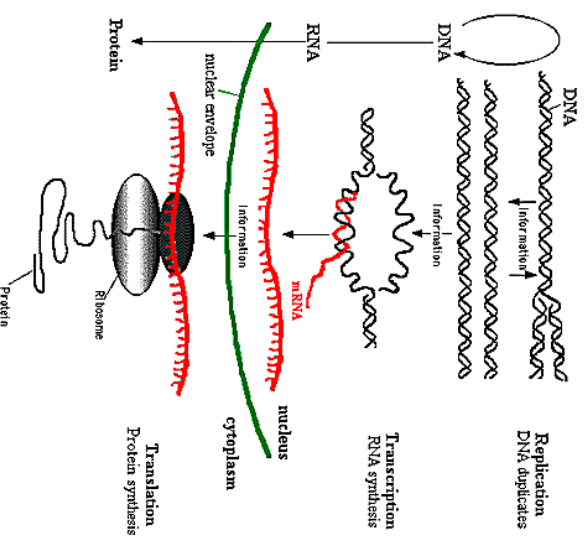
What is BLAST?

- **B**asic **L**ocal **A**lignment **S**earch **T**ool, or BLAST, is an algorithm for comparing biological sequence information.
- It can handle both amino acid sequences (encoding proteins) or nucleotide sequences (from DNA).
- BLAST is one of the most widely used bioinformatics programs.
- It is actually a suite of programs for different applications.

3

The Central Dogma of Molecular Biology

- DNA remains in the nucleus.
- mRNA carries information from DNA and serve as templates that order different proteins.
- Proteins are responsible for cell function.

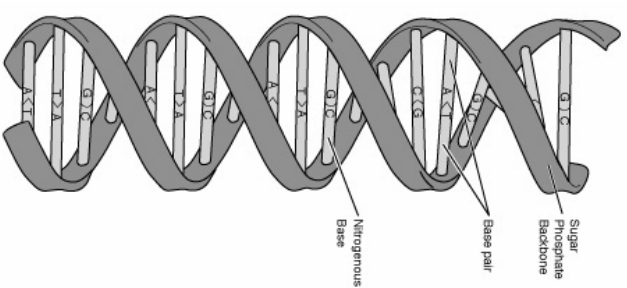


The Central Dogma of Molecular Biology

4

DNA/Nucleotide Sequences

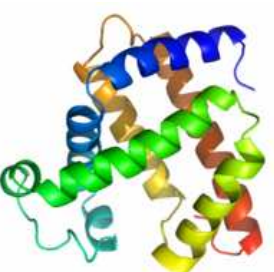
- DNA contains the genetic instructions used in the development and functioning of all living organisms.
- A nucleotide sequence consists of the letters *A*, *C*, *G*, and *T*, representing the four nucleotide subunits of a DNA strand - adenine, cytosine, guanine and thymine.
- Base Pairing: A bonds only to T, and C bonds only to G



5

Proteins / Amino Acid Sequences

- Proteins are responsible for cell function.
- Amino acids are the building blocks of proteins.
- There are 20 amino acids: A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V



A representation of the 3D structure of myoglobin.

6

Who developed BLAST?

- Developed by Stephen Altschul, Warren Gish, David Lipman at the U.S. National Center for Biotechnology Information (NCBI), Webb Miller at Pennsylvania State University, and Gene Myers at the University of Arizona.
- The original paper “Basic local alignment search tool” appeared in the Journal of Molecular Biology in 1990.

7

Where can BLAST be found?

- The BLAST program can be accessed for free over the web or downloaded from NCBI.
- At CSU, BLAST can be accessed through the CBC cluster for faster running time.

8

Web Access: www.ncbi.nlm.nih.gov

The screenshot shows the NCBI homepage. At the top, there's a navigation bar with 'All Databases' and 'BLAST' (circled in red) highlighted. Below the navigation bar is a search bar with a 'Go' button. The main content area is divided into several sections: 'What does NCBI do?' (a blue box with a right-pointing arrow), '100 Gigabases' (a yellow box with a blue border), and 'Hot Spots' (a dark blue box with a right-pointing arrow). The 'Hot Spots' section contains several links: Assembly Archive, Clusters of orthologous groups, Coffee Break, Genes & Disease, NCB Handbook, Electronic PCR, Entrez Home, Entrez Tools, Gene expression omnibus (GEO), Human genome resources, Malaria genetics & genomics, and Map Viewer. The '100 Gigabases' section contains text about the European Molecular Biology Laboratory and the DNA Databank of Japan. The 'What does NCBI do?' section contains text about the center's history and mission.

9

How does BLAST work?

- BLAST searches for similar subsequences between a query (or target) sequence and a sequence database:
 1. Scan database for exact matches of a small fixed length between the query and sequences database.
 2. Initiate extensions from these matches.
 3. Performs a gapped alignment between the query sequence and the database sequence using a variation of the Smith-Waterman algorithm.
 4. Statistically significant alignments are then displayed to the user.

10

Scan Database... Initiate Extensions

Protein BLAST requires two hits

GTQITVEDLIFYNI

<----- TVE FYN ----->

two neighborhood words

Nucleotide BLAST requires exact matches

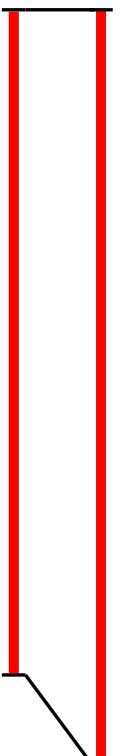
ATGCCATGCTTAATTGGCTT

<----- CATGCTTAATT ----->

exact word match

Global vs Local Alignment

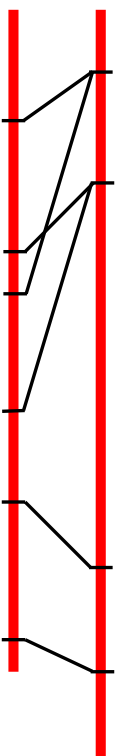
Seq 1



↓ Global alignment

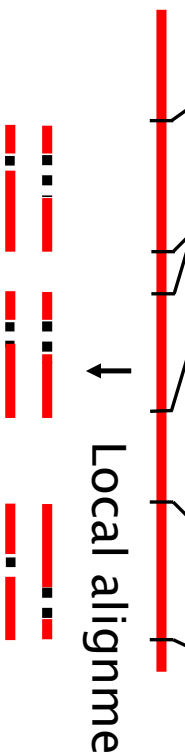


Seq 1



↓ Local alignment

Seq 2



Why is BLAST used?

- Infer functional and evolutionary relationships between sequences
- Identify members of gene families
- Look for similar genes across species
- Select primers for PCR
- Annotate probes from a microarray experiment
- Search for SNPs
- Align sequences

13

Example 1: Cross Species Hybridization

- Campus researchers were interested in studying gene expression in common bean (*Phaseolus vulgaris*).
- However, they used the Affymetrix Soybean Genome Array because there is no phaseolus array available from Affymetrix.



14

Query Sequence

- Suppose that probe set Gma.2019.1.S1_at was identified as representing a gene of interest.
- From the Affymetrix website we can find the target sequence:

```
CCAAGTTGAGACCCTTATTGACATTTCAGTGATAGATATTGTCAA  
GGCATGCTTCTGATGAGGTCTACCTTGGCCAAAGGATTAATCCAATTT  
GGACTACGGATTTCAAAAGGCATTTGGAAGCTTTCAAAAAAGTTTGGAAACA  
AACTGGCAGAAATTGAGGGAAAAATCACACAGAGGAACAATGATCCAA  
GTCTGAAAAAGCCGACATGGCCAGTTTCAGCTTCCATACACATTTGCTCC  
ATCGTTCAAGTGAGGAAGGATGAGTTTCAAGAATTCGCCAACAGTA  
TCCTCATCTAAATGTGTGTGTTTGGCTTATCTATTGTG
```

- This sequence represents the gene for lipoxxygenase in soybean. But does it match to any known genes in Phaselous?

15

Nucleotide BLAST Search

- Enter Query Sequence: simply paste our query sequence into the window.
- Choose Search Set:
 - Database: Nucleotide Collection (nr/nt)
 - Organism: Phaseolus vulgaris (taxid:3885)
- Program Selection/Algorithm:
 - Megablast (highly similar sequences)
 - Discontiguous mega blast (more dissimilar sequences)
- ▫ **Blanstn** (somewhat similar sequences)

16

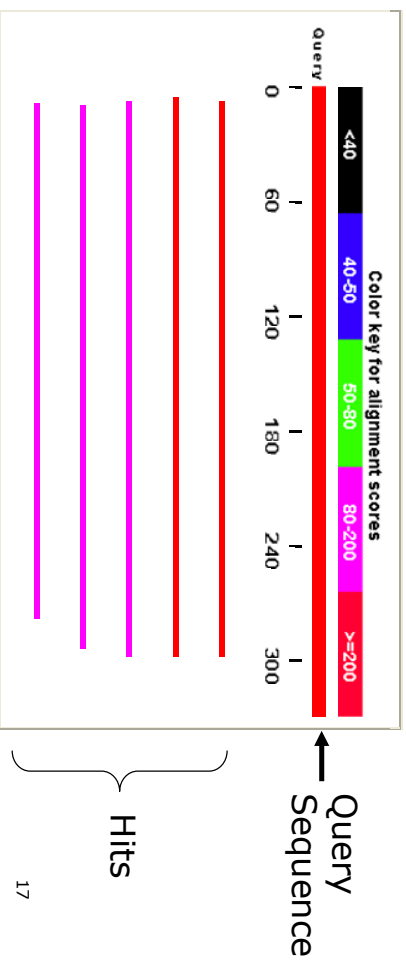
BLAST Results

Your search is limited to records matching entrez query: txid3885 [ORGN].

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, environmental samples or phase 0, 1 or 2 HTGS sequences)
 5,298,904 sequences; 20,720,589,268 total letters

Query: Length=329

Distribution of Blast Hits on the Query Sequence



BLAST Results

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
U76687.2	Phaseolus vulgaris lipoxygenase mRNA, complete cds	232	232	88%	7.00E-62	77%
X63525.1	P. vulgaris loxA gene for lipoxygenase	219	219	90%	4.00E-58	76%
AF234983.1	Phaseolus vulgaris lipoxygenase gene, complete cds	194	194	88%	2.00E-50	74%
AF204210.2	Phaseolus vulgaris lipoxygenase (LOX4) mRNA, partial cds	181	181	87%	1.00E-46	73%
X63521.1	P. vulgaris mRNA for lipoxygenase	170	170	82%	2.00E-43	73%

(Brief) Interpretation of Results

- Accession: Sequence ID
- Max Score/Total Score: Score assigned by BLAST. In general the higher the score, the better the match between the query sequence and the sequence found in the database.
- Bit Score: Standardized Score
- Query Coverage: The percentage of the query sequence covered by each match.
- E value: The E value is the number the number of hits one can "expect" to see just by chance when searching a database of a particular size.
- Max ident: Percentage of exact matches

19

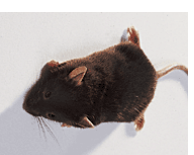
Top Match

```
> gb|U76687.2|PVU76687 Phaseolus vulgaris lipoxxygenase mRNA, complete cd
Length=3038
Score = 232 bits (256), Expect = 7e-62
Identities = 226/291 (77%), Gaps = 0/291 (0%) strand=Plus/Plus
Query 9 GAGACCCCTTATTGACATTTCAGTGATAGAGATATTGTCAAGGCATGCTTGTGATGAGGTC
||||| ||| | ||| ||||| ||||||| ||||||| ||||||| ||||||| |||||||
Sbjct 2581 GAGGCCATTGTGGACCTTCTGTGATAGAGATATTGTCAAGACATGCTTGTGATGAGGTC
Query 69 TACCTTGGCCAAAGGGATTAATCCAAATTGGACTACGGATTCCAAGGCATTTGGAAGCTTTC
|| ||||| || ||||| ||||| ||||||| || || ||||| || ||||| ||
Sbjct 2641 TATCTTGGACAGAGGGACAATCCTTAATTGGACTGTGATATACAAAGGCTCTTCAAGGCTTTT
Query 129 AAAAAAGTTTGGAAAACAACAACTGGCAGAAATTGAGGGAAAAATCACACAGAGGAACAATGAT
||||| ||||||| ||||||| || ||||| || ||||| || ||
Sbjct 2701 CAAAAAGTTTGGAAAACAACAACTGGAAGAAATTGAGAA TAAGATCTTAGGAAGGAACAACAAT
Query 189 CCAAGTCTGAAAAGCCCGACATGGGCCAGTT CAGCTTCCATAACACATTTGCTCGATGTTCA
||||| || || ||| ||||||| || || ||||| ||||| || || ||
Sbjct 2761 TCAAGTCTCAGAAAACCGTGTGGGCCAGTTAAGATGCCCTACACTGTGCTTCTTCCCTAAC
Query 249 AGTGAAGGAA GGGATGAGTTCAAAAGGAATTC CCAACAGTATCTCCATCTAA
|||| ||||| || | ||||| ||||| ||||||| ||||| || |||
Sbjct 2821 AGTAAGGAAGGTCACACTTTCAGAGGAATCCCAACAGCATCTCTATTAA 2871
```

20

Example 2: Leptin

- An article (“Positional cloning of the mouse obese gene and its human homologue”) published in Nature in 1994 announced the discovery of the *obese* gene.
- Jeffrey Friedman and colleagues at Rockefeller University discovered that mice become fat when they lacked a single gene, *ob*.
- Friedman named the missing gene's product leptin, after the Greek word for thin (*leptos*).
- Finding this mouse gene enabled researchers to discover a homologous gene in humans.



23

Finding the *ob* Gene

- The authors located the approximate region of the gene, then located genes in the region of interest.
- The *ob* gene was located and sequenced.
- “*ob* encodes a 4.5-kilobase (kb) adipose tissue messenger RNA with a highly conserved 167-amino-acid open reading frame.”

24

Query Sequence

- We will use the 167-amino-acid sequence from the paper:

```
MCWRPLGRFLMWSYLSYVOAVPIQKVQDDTKTLIKTIVTRINDISHTOS
VSAKQRYVTGLDFIPGLHPILSLSKMDQTLAVYVYQQVLTSLPSQNVLQIAND
LENLRDLHLHLAFSKSCSLPQTSGLQKPESLDGVLEASLXSTEVVALSRL
QGSLQDILQQILDVSP EC
```

- At the time of publication, “A database search of the *ob* protein using the BLAST program identified no significant homology to any sequences in GenBank.”

25

Protein BLAST Search

- Enter Query Sequence: simply paste our query sequence into the window.
- Choose Search Set/Database:
 - Non-redundant protein sequences (nr)
 - Reference proteins (refseq_protein)
 - Swissprot protein sequences (swissprot)
 - Others
- Program Selection/Algorithm:
 - ■ blastp(protein-protein BLAST)
 - PSI-BLAST (Position-Specific Iterated BLAST)
 - PHI-BLAST (Pattern Hit Initiated BLAST)

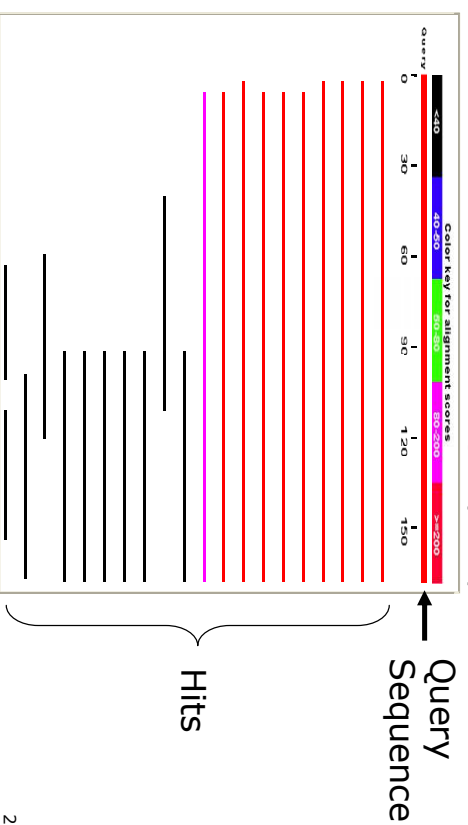
26

BLAST Results

Database: NCBI Protein Reference Sequences
 3,168,478 sequences; 1,148,749,149 total letters

Query: Length=167

Distribution of Blast Hits on the Query Sequence



27

BLAST Results

Sequences producing significant alignments:

ref	NP_032519.1	leptin [Mus musculus]	<u>298</u>	5e-80
ref	NP_037208.1	leptin [Rattus norvegicus]	<u>293</u>	2e-78
ref	NP_000221.1	leptin precursor [Homo sapiens]	<u>257</u>	1e-67
ref	XP_519353.2	PREDICTED: leptin [Pan troglodytes]	<u>256</u>	3e-67
ref	NP_001009850.1	leptin [Felis catus]	<u>252</u>	4e-66
ref	NP_999005.1	leptin [Sus scrofa]	<u>251</u>	6e-66
ref	NP_776353.2	leptin [Bos taurus]	<u>248</u>	8e-65
ref	NP_001036220.1	leptin [Macaca mulatta]	<u>247</u>	1e-64
ref	NP_001003070.1	leptin [Canis familiaris]	<u>243</u>	2e-63



RefSeq Accession Numbers

Sequence Descriptions

28

Top Match

>ref|NP_032519.1| leptin [Mus musculus]
Length=167

Score = 298 bits (763), Expect = 5e-80, Method: Composition-based stats.
Identities = 165/167 (98%), Positives = 165/167 (98%), Gaps = 0/167 (0%)

```
Query 1 MWRPILGRFLWMSYLSYVOAVPIQKVQDDPKTLIKTIYTRINDISHTOSVSAKQRYTGL 60
      MWRPILGRFLWMSYLSYV AVPIQKVQDDPKTLIKTIYTRINDISHT SVSAKQRYTGL
Sbjct 1 MWRPILGRFLWMSYLSYVQAVPIQKVQDDPKTLIKTIYTRINDISHTQSVSABQRYTGL 60

Query 61 DFIPGLHPILSLSKMDQTLAVYQQVLTSLPQNVLQIANDLENLRDLHLHLAFSKSGSLP 120
      DFIPGLHPILSLSKMDQTLAVYQQVLTSLPQNVLQIANDLENLRDLHLHLAFSKSGSLP
Sbjct 61 DFIPGLHPILSLSKMDQTLAVYQQVLTSLPQNVLQIANDLENLRDLHLHLAFSKSGSLP 120

Query 121 QTSGLQKPESLDGVEEASLSTEVVALSRLQGSLLQDILQQLDVSPEC 167
      QTSGLQKPESLDGVEEASLSTEVVALSRLQGSLLQDILQQLDVSPEC
Sbjct 121 QTSGLQKPESLDGVEEASLSTEVVALSRLQGSLLQDILQQLDVSPEC 167
      QTSGLQKPESLDGVEEASLSTEVVALSRLQGSLLQDILQQLDVSPEC 29
```

Second Match

>ref|NP_037208.1| leptin [Rattus norvegicus]
Length=167

Score = 293 bits (749), Expect = 2e-78, Method: Composition-based stats.
Identities = 159/167 (95%), Positives = 163/167 (97%), Gaps = 0/167 (0%)

```
Query 1 MWRPILGRFLWMSYLSYVOAVPIQKVQDDPKTLIKTIYTRINDISHTOSVSAKQRYTGL 60
      MWRPILGRFLWMSYLSYV AVPI KVQDDPKTLIKTIYTRINDISHT SVSA+DRVTGL
Sbjct 1 MWRPILGRFLWMSYLSYVQAVPIHKVQDDPKTLIKTIYTRINDISHTQSVSABQRYTGL 60

Query 61 DFIPGLHPILSLSKMDQTLAVYQQVLTSLPQNVLQIANDLENLRDLHLHLAFSKSGSLP 120
      DFIPGLHPILSLSKMDQTLAVYQQ+LTSLPQNVLQIA+DENLRDLHLHLAFSKSGSLP
      DFIPGLHPILSLSKMDQTLAVYQQ+LTSLPQNVLQIA+DENLRDLHLHLAFSKSGSLP
Sbjct 61 DFIPGLHPILSLSKMDQTLAVYQQVLTSLPQNVLQIADLENLRDLHLHLAFSKSGSLP 120

Query 121 QTSGLQKPESLDGVEEASLSTEVVALSRLQGSLLQDILQQLDVSPEC 167
      QTSGLQKPESLDGVEEASLSTEVVALSRLQGSLLQDILQQL+SPEC
      QT GLQKPESLDGVEEASLSTEVVALSRLQGSLLQDILQQL+SPEC
Sbjct 121 QTRGLQKPESLDGVEEASLSTEVVALSRLQGSLLQDILQQLDVSPEC 167
      QTRGLQKPESLDGVEEASLSTEVVALSRLQGSLLQDILQQLDVSPEC 30
```

Third Match

```
>ref|NP_000221.1| leptin precursor [Homo sapiens]
Length=167
Score = 257 bits (656), Expect = 1e-67, Method: Composition-based stats.
Identities = 137/167 (82%), Positives = 150/167 (89%), Gaps = 0/167 (0%)

Query 1 MWRPILGRFLMWSYLSYVOAVPIQKVQDDPKTLIKTIYTRINDISHTOSVSAKQRYTGL 60
      M W LC FLMLW YL YV AVPIQKVQDDPKTLIKTIYTRINDISHT
Sbjct 1 MHWGITLGGFLWMPYLFYVQAVPIQKVQDDPKTLIKTIYTRINDISHTQSVSSKQKYTGL 60

Query 61 DFIPGLHPILSLSKMDQTLAVYQOVLTSLSQNVLIQIANDLENLRDLHLHLAFSKSGSLP 120
      DFIPGLHPIL+LSKMDQTLAVYQO+LTS+PS+NV+QI+NDLENLRDLHL+LAFSKSC LP
Sbjct 61 DFIPGLHPILTLTKMDQTLAVYQOILLTSMPSRNVIQISNDLENLRDLHLVLAFAFSKCHLP 120

Query 121 QTSGLQKPESLDGVLEASLYSTEVAALSRLOGSLQDILLQQLDVSPEC 167
      SGL+ +SL GVLEAS YSTEVAALSRLOGSLQD+L QLD+SP C
Sbjct 121 WASGLETLIDSLGGVLEASGYSTEVAALSRLOGSLQDMLWQLDLSPGC 167

Sbjct 121 WASGLETLIDSLGGVLEASGYSTEVAALSRLOGSLQDMLWQLDLSPGC 31
```

More Distant Match

```
>ref|NP_878565.1| exonuclease V, beta chain [Candidatus Blochmannia floridanus]
Length=1208
Score = 35.0 bits (79), Expect = 0.95, Method: Composition-based stats.
Identities = 24/76 (31%), Positives = 36/76 (47%), Gaps = 5/76 (6%)

Query 41 RINDISHTOSVSAKQRYTGLDFIPGLHPILSLSKMDQTLAVYQOVLTSLSQNVLIQIAN 99
      R ND H +S + GL+F P +S + + Q++L S P N+LQ ++
Sbjct 734 RFNDDHHLIKISTIHQSKGLEFPITTYLPIFCYTTNTRKRFNNQKILLQSTPPYANLLQSSST 793

Query 100 -----DLENLRDLHLHL 111
      IDLE L + L LL
Sbjct 794 LDFFDLERRLSEDLRL 809
```


Example 2 Conclusions

- ❑ Homologs to leptin have been identified in many other species besides mice.

35

Options for Advanced Blasting: Nucleotide

- ❑ Many of the algorithm parameters can be adjusted:
 - Word Size
 - Scoring System and Gap Costs
 - Expect Threshold
 - Filtering: Filtering can eliminate statistically significant but biologically uninteresting reports from the blast output
 - ❑ Low Complexity Regions (Ex: AAATTAATAATAAT)
 - ❑ Species Specific Repeats

36

Options for Advanced Blasting: Protein

- ❑ Many of the algorithm parameters can be adjusted:
 - Word Size
 - Substitution Matrices:
 - PAM30, PAM70
 - BLOSUM45, BLOSUM62, BLOSUM80
 - Gap Costs
 - Expect Threshold
 - Filtering

37

Other BLAST Programs

- ❑ Megablast (for Nucleotide sequences) is intended for comparing a query to closely related sequences but is very fast.
- ❑ PSI (Position-Specific Iterated) BLAST (for Protein sequences) can be used to find distant relatives of a protein.
- ❑ mpBLAST segments the BLAST database and distributes it across cluster nodes.

38

Specialized BLAST Programs

- ❑ Search for SNPs (Single Nucleotide Polymorphisms). A SNP is a DNA sequence variation occurring when a single nucleotide differs between members of a species (or between paired chromosomes in an individual)
- ❑ Find sequences with gene expression profiles.
- ❑ Align two sequences.
- ❑ Find conserved domains in a sequence.

39

Alternatives to BLAST

- ❑ BLAT (**B**LAST **L**ike **A**lignment **T**ool) is much faster than BLAST, but less sensitive.
- ❑ BLASTZ is a multiple sequence alignment program.
- ❑ PatternHunter is commercial homology search software that is significantly faster than BLAST.
- ❑ HMMER uses hidden Markov models to do sensitive database searching.
- ❑ FASTA is a DNA and protein sequence alignment software package.

40