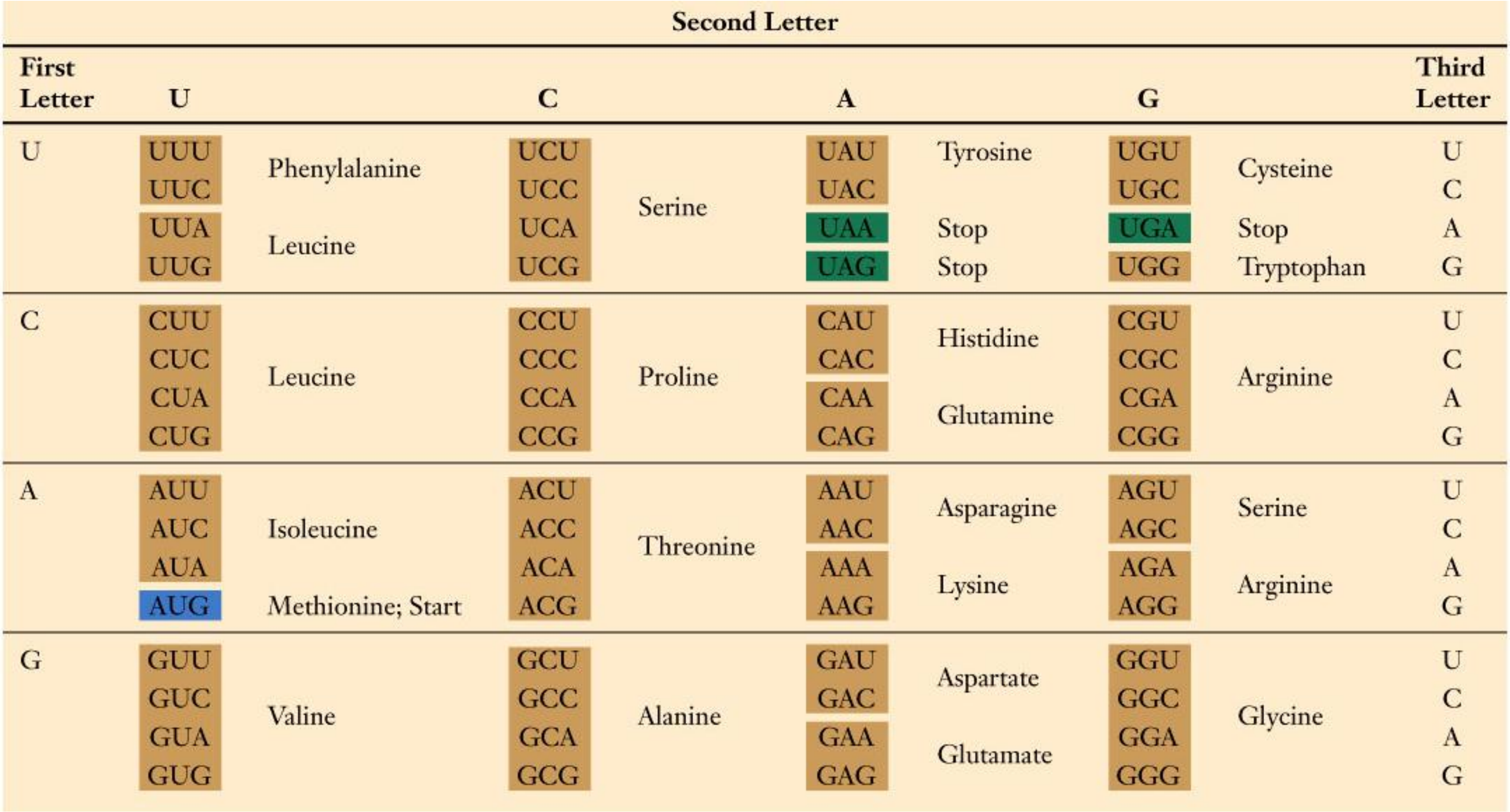
Lecture 7: Sequencing a Cloned Gene – Analysis and Annotation

Universal Genetic Code



	Second Letter								
First Letter	U		C		A		G		Third Letter
U	UUU	Phenylalanine	UCU	Serine	UAU	Tyrosine	UGU	Cysteine	U
	UUC		UCC		UAC		UGC		C
	UUA	Leucine	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Tryptophan	G
C	CUU	Leucine	CCU	Proline	CAU	Histidine	CGU	Arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Glutamine	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Isoleucine	ACU	Threonine	AAU	Asparagine	AGU	Serine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lysine	AGA	Arginine	A
	AUG	Methionine; Start	ACG		AAG		AGG		G
G	GUU	Valine	GCU	Alanine	GAU	Aspartate	GGU	Glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glutamate	GGA		A
	GUG		GCG		GAG		GGG		G

Test Sequences

btr-3 Nucleotide Sequence

ATGAAAGTGGAGAGTTGGTTGCACTTGGGTTGGTTGCTGGGGTTGCTGCTGGTCCTGTTGCCGTTGGTC
CGATGCCAAGGATGGGGCGAACCACGGTTCGAGACGGGAAATGTGGAAAATATATCACTCGCCGCATAC
AACGAGGCGCAGCTGCAGCAAGATGTCTGGATGGTGGAGGAGATGGATGCACCGTTCGTGCTGCTCTAC
ATCAATTACCAAGGACCGTCCGAGCCTACGATACGCGAGTCACCGGCCGATCTTGACGCAAGGCTACAGC
TGTCCGAGGCTGGCCGCTGGTCGATCGTAATCAATCGCCGTCAGGACTACGAGGTGCATCAGCGTAGCA
GTCTCATTCTGCTGGCCGTCGAATCCACGGCTATCCCGTACGCGATCGTGGTCAACTTGGTGAACGTGCTG
GACAATGCGCCCGTCATGACGGCCCAAGGTAGCTGTGAGATTGAGGAGTTGCGCGGGGACTTTGTGAC
GGACTGTCTGTTTAACGTGTACCATGCGGACGGGTTCGAGGAGAATGGCATTGGCAATTCGAGCACGAA
CGAGCTGTCGTTCGAGATCGGTGATGTGGCCGGTGCGCGGGACCACTTTACGTACGTGCCCTCCACGGT
GACCCCTTCCCAGCCGATCTACAACAAGCTGTTCAATTTGAAAGTTTTAAAGCAGCTGGACTACACCGAG
AACGCTATATTTAACTTCATCACCACCGTGTACGACCTAGACCGGACGCACTCCTTCAAGATGAGTACGAT
CGTTCAGGTNCGCAACGTCGATAGCCGGCCTCCGATCTTTAGCCGACCGTTCNCCAGCGAACGNATCATG
NAAANGgAANCATTTTACGCgANCGTGATCGCANtCGAcCGTGACACTgGaCTAAACAAACCGATCTGTT
ACGAGCTGACGGCTCTAGTACCGGAATATCAGAAATATTTCGATATTGGACAAACTGATGGAAAGCTGAC
CGTGCACCCGATTGATCGAGATGCGG

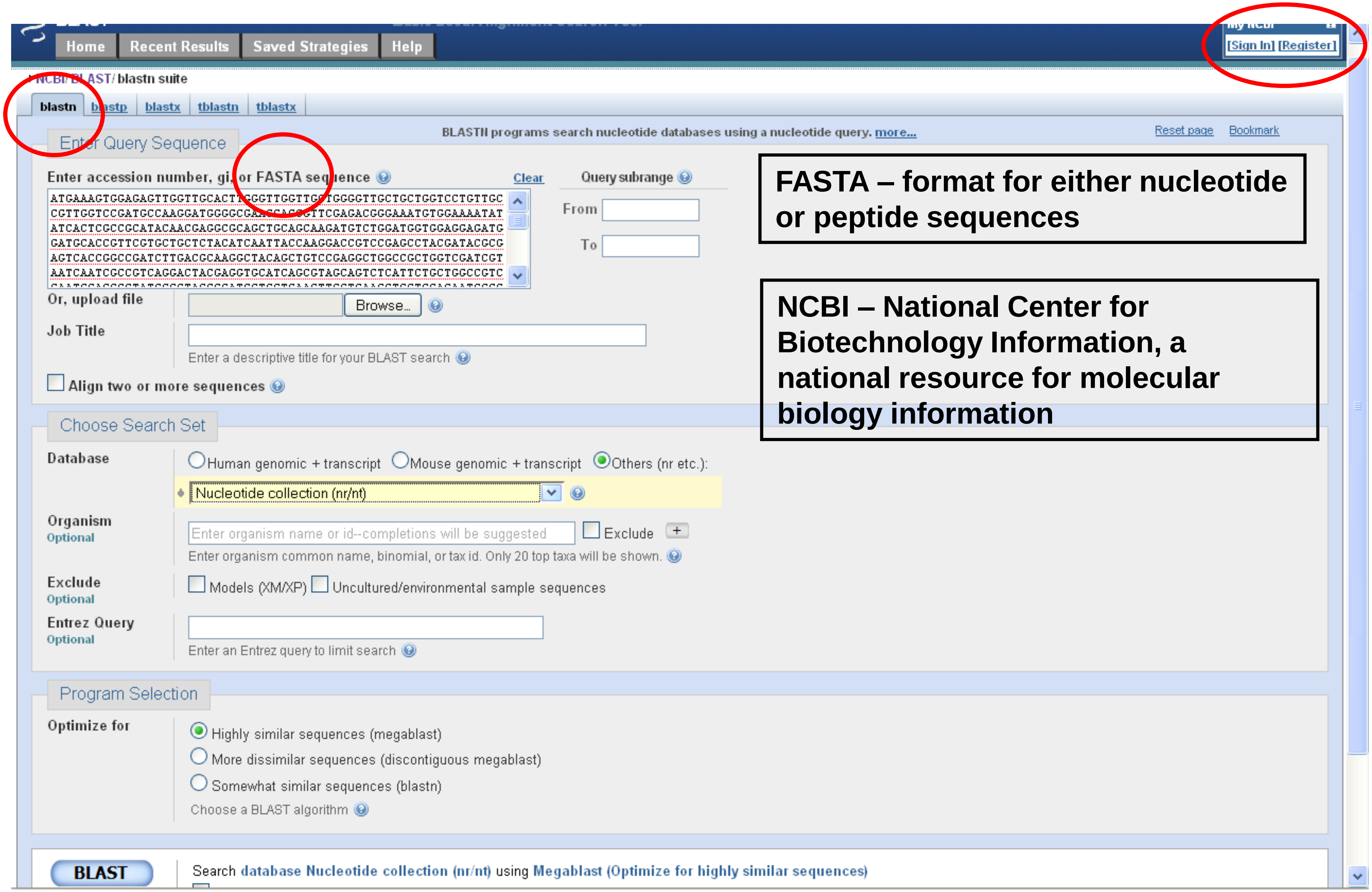
BTR-3 Predicted Protein Sequence

MKVESWLHLGWLLGLLLVLLPLVRCQGWGEPRFETGNVENISLAAYNEAQLQQDVWMVEEMDAPFVLLYI
NYQGPSEPTIRESPADLDARLQLSEAGRWSIVINRRQDYEVHQRSSLILLAVESTAIPYAIVVNLVNVLDNAPV
MTAQGSCEIEELRGDFVTDCLFNVYHADGFEENGIGNSSTNELSFEIGDVAGARDHFTYVPSTVTPSQPIYN
KLFNLKVLKQLDYTENAIFNFITTVYDLDRTHSFKMSTIVQVRNVDSRPPIFSRPFXSERIMXXEXFYAXVIAX
DRDTGLNKPICYELTALVPEYQKYFDIGQTDGKLTVHPIDRDA

National Center for Biotechnology Information (NCBI) BLAST Family of Programs

- **Blastp – compares an amino acid query sequence against a protein sequence database; recognizes evolutionary conservation**
- **Blastn – compares a nucleotide query sequence against a nucleotide sequence database**
- **Blastx – compares a nucleotide query sequence translated in all reading frames against a protein sequence database**
- **Tblastn – compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames**
- **Tblastx - compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database**
- **BLAST Line (BLink)** – **A link option on protein records that displays the results of a pre-computed BLAST search of that protein against all other protein sequences at NCBI**

Entering the Nucleotide Sequence



Enter Query Sequence

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter accession number, gi, or FASTA sequence

```
ATGAAAGTGGAGAGTTGGTTGCACTTGGGTTGGTTGGTGGGGTTGCTGCTGGTCCTGTTGC
CGTTGGTCCGATGCCAAGGATGGGGCGAACGACGGTTCGAGACGGAAATGTGGAAAATAT
ATCACTCGCCGCATACAACGAGGCGCAGCTGCAGCAAGATGTCTGGATGGTGGAGGAGATG
GATGCACCGTTCGTGCTGCTCTACATCAATTACCAAGGACCGTCCGAGCCTACGATACGCG
AGTCACCGGCCGATCTTGACGCAAGGCTACAGCTGTCCGAGGCTGGCCGCTGGTCGATCGT
AATCAATCGCCGTCAGGACTACGAGGTGCATCAGCGTAGCAGTCTCATTCTGCTGGCCGTC
```

Query subrange

From

To

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.):

Nucleotide collection (nr/nt)

Organism
Optional

Enter organism name or id--completions will be suggested □ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude
Optional

□ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query
Optional

Enter an Entrez query to limit search

Program Selection

Optimize for

● Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

FASTA – format for either nucleotide or peptide sequences

NCBI – National Center for Biotechnology Information, a national resource for molecular biology information

Important Terms Used in BLAST

- Accession number: unique identification number for the sequence

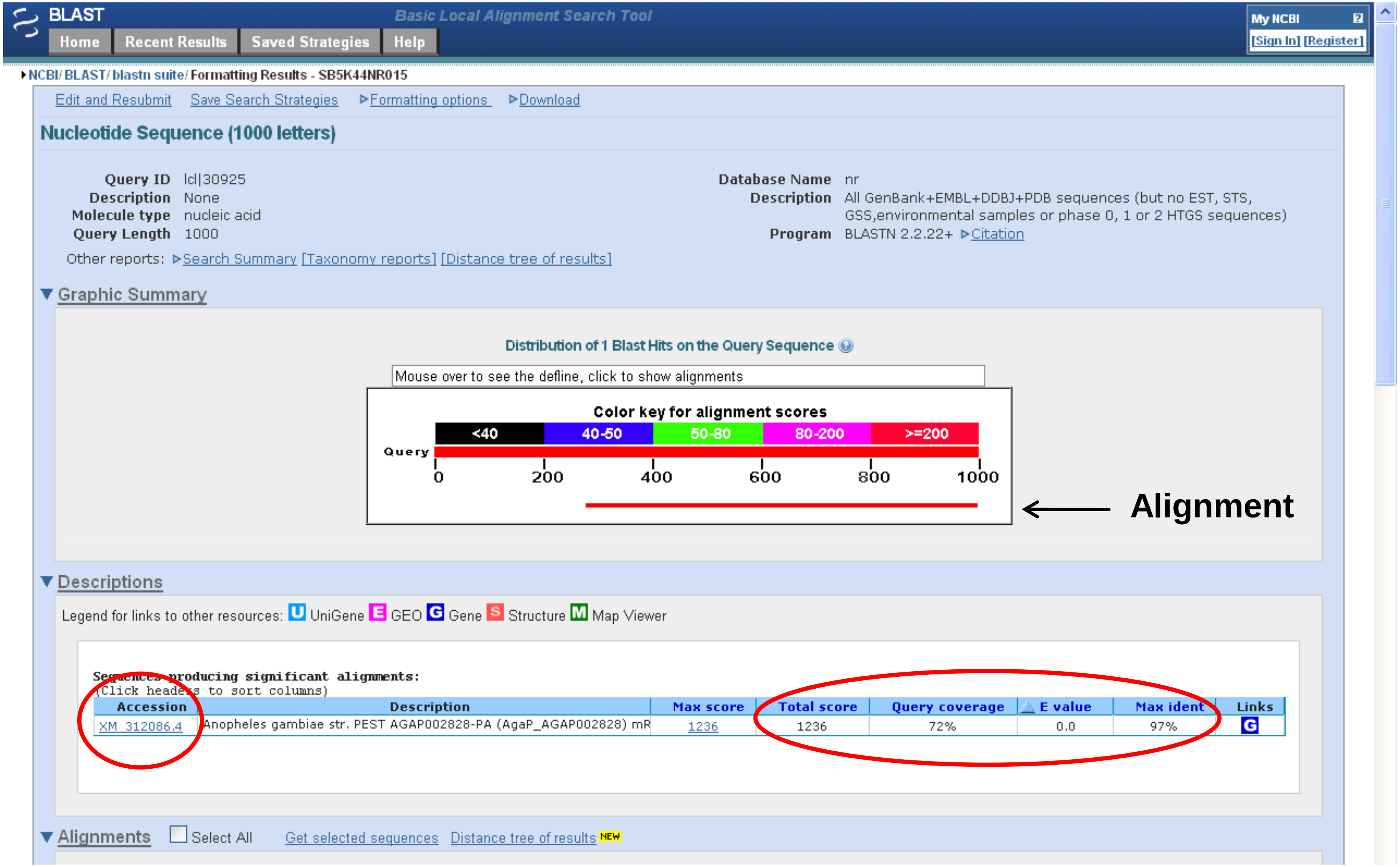
- Score: score calculated for this particular match, using a scoring matrix

- Query coverage: percentage of the query sequence that matched the sequence in the database. (In this case, nucleotide 280 to nucleotide 1000 = 72% of the query sequence)

- E value (Expect value): the number of hits one can "expect" to see by chance when searching a database of a particular size. (In the following example, 0% implies that it is a significant match)

- Maximum identity: percentage of nucleotides matched being identical (In the following example, 706 of the 727 nucleotides matched were identical to each other, which is 97%)

Results Obtained



Nucleotide Sequence (1000 letters)

Query ID lcl|30925
Description None
Molecule type nucleic acid
Query Length 1000

Database Name nr
Description All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
Program BLASTN 2.2.22+ Citation

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

Graphic Summary

Distribution of 1 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments

Alignment

Descriptions

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

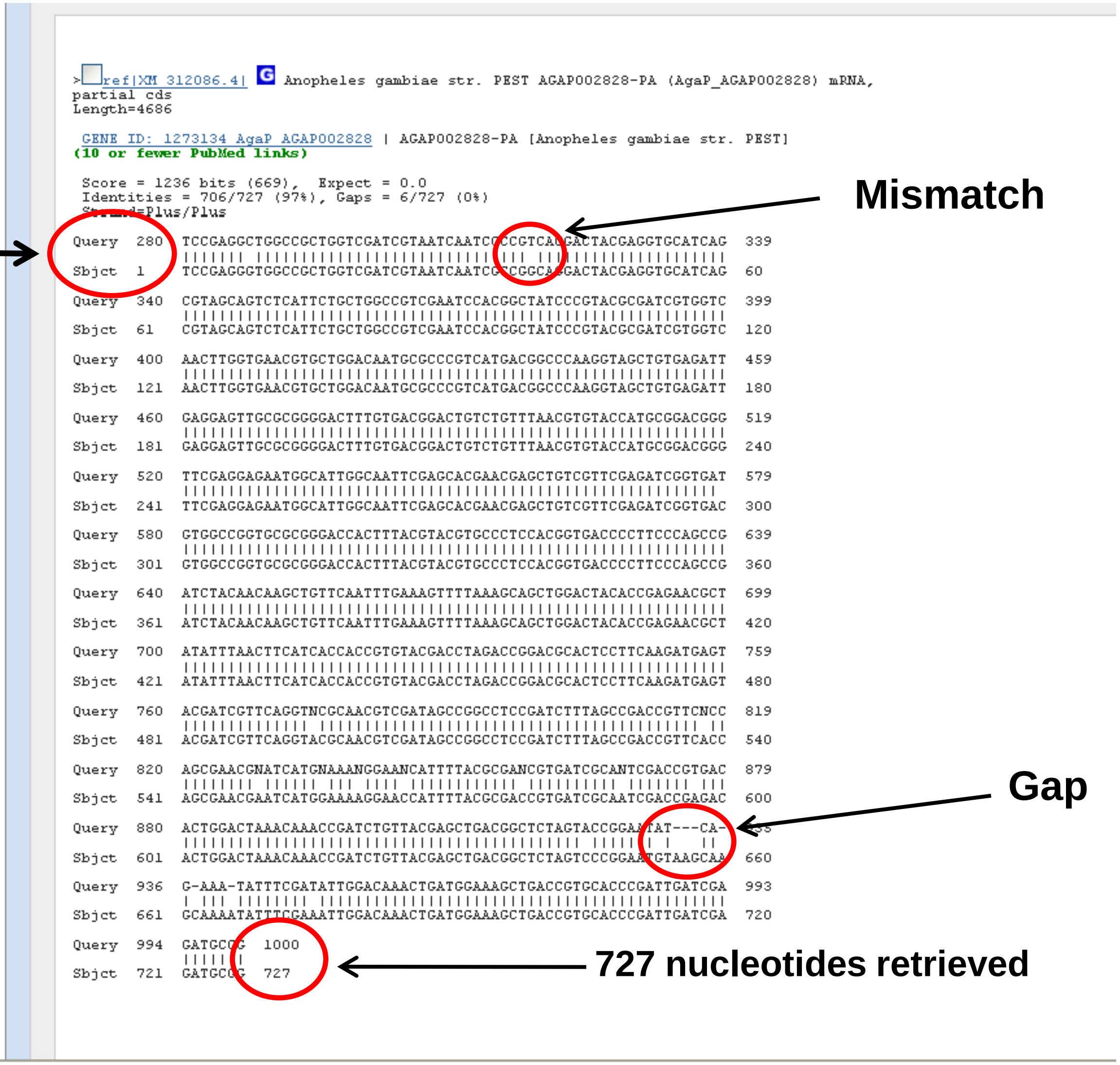
Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
XM_312086.4	Anopheles gambiae str. PEST AGAP002828-PA (AgaP_AGAP002828) mR	1236	1236	72%	0.0	97%	G

Alignments □ Select All Get selected sequences Distance tree of results NEW

Sequence Comparison

Nucleotide 280 of query matched nucleotide 1 of the database sequence



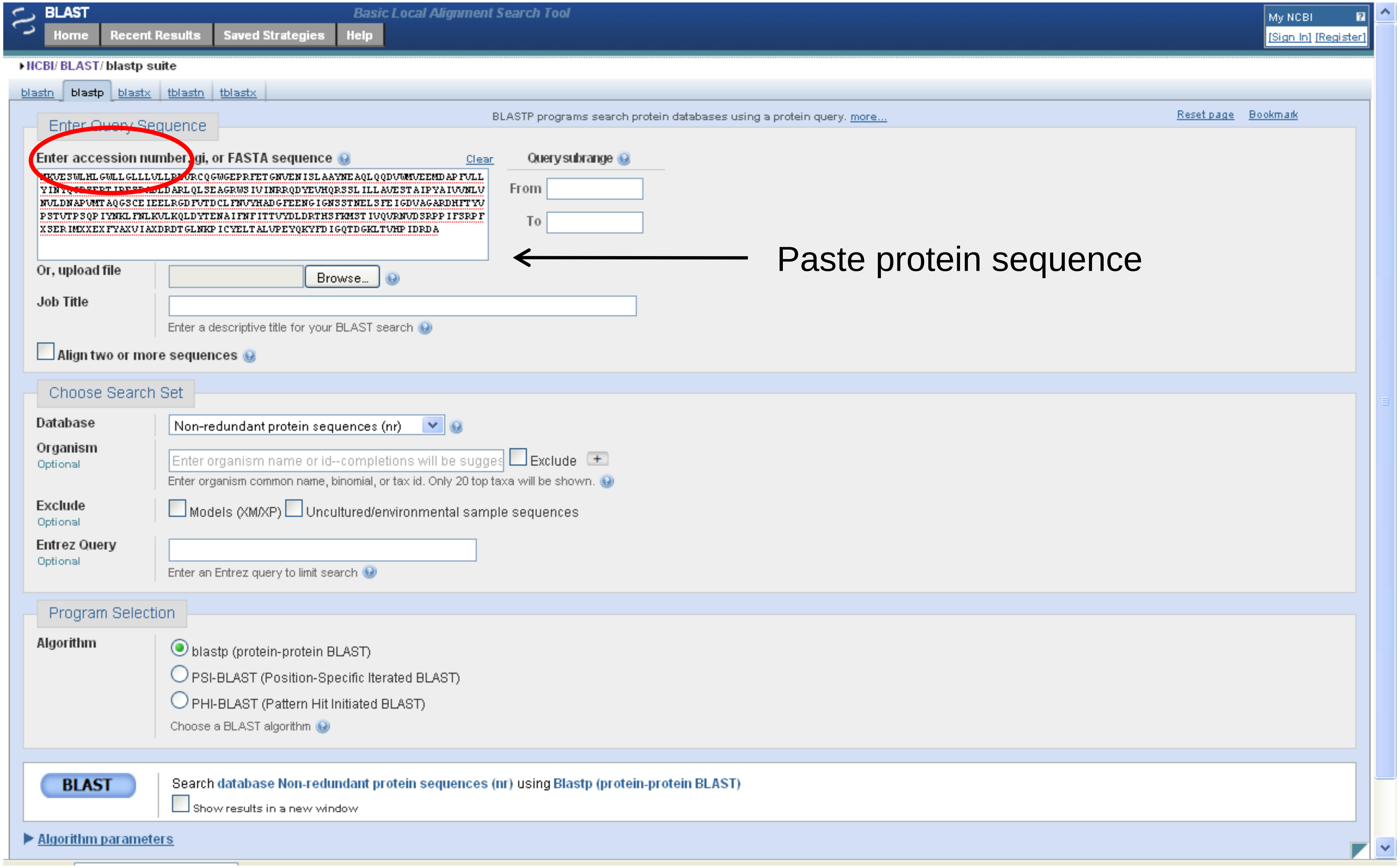
```
>ref|XM 312086.4| G Anopheles gambiae str. PEST AGAP002828-PA (AgaP_AGAP002828) mRNA,
partial cds
Length=4686

 GENE ID: 1273134 AgaP AGAP002828 | AGAP002828-PA [Anopheles gambiae str. PEST]
(10 or fewer PubMed links)

 Score = 1236 bits (669),  Expect = 0.0
 Identities = 706/727 (97%), Gaps = 6/727 (0%)
 Strand=Plus/Plus

Query  280   TCCGAGGCTGGCCGCTGGTCGATCGTAATCAATCGCCGTCACGACTACGAGGTGCATCAG  339
             ||||||| |||||||||||||||||||||||||||||| |||||||||||||||||||||
Sbjct  1     TCCGAGGGTGGCCGCTGGTCGATCGTAATCAATCGCCGGCACGACTACGAGGTGCATCAG  60

Query  340   CGTAGCAGTCTCATTCTGCTGGCCGTCGAATCCACGGCTATCCCGTACGCGATCGTGGTC  399
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    CGTAGCAGTCTCATTCTGCTGGCCGTCGAATCCACGGCTATCCCGTACGCGATCGTGGTC  120

Query  400   AACTTGGTGAACGTGCTGGACAATGCGCCCGTCATGACGGCCCAAGGTAGCTGTGAGATT  459
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   AACTTGGTGAACGTGCTGGACAATGCGCCCGTCATGACGGCCCAAGGTAGCTGTGAGATT  180

Query  460   GAGGAGTTGCGCGGGGACTTTGTGACGGACTGTCTGTTTAACGTGTACCATGCGGACGGG  519
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181   GAGGAGTTGCGCGGGGACTTTGTGACGGACTGTCTGTTTAACGTGTACCATGCGGACGGG  240

Query  520   TTCGAGGAGAATGGCATTGGCAATTCGAGCACGAACGAGCTGTCGTTCGAGATCGGTGAT  579
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| 
Sbjct  241   TTCGAGGAGAATGGCATTGGCAATTCGAGCACGAACGAGCTGTCGTTCGAGATCGGTGAC  300

Query  580   GTGGCCGGTGCGCGGACCACTTTACGTACGTGCCCTCCACGGTGACCCCTTCCCAGCCG  639
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301   GTGGCCGGTGCGCGGACCACTTTACGTACGTGCCCTCCACGGTGACCCCTTCCCAGCCG  360

Query  640   ATCTACAACAAGCTGTTCAATTTGAAAGTTTTAAAGCAGCTGGACTACACCGAGAACGCT  699
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361   ATCTACAACAAGCTGTTCAATTTGAAAGTTTTAAAGCAGCTGGACTACACCGAGAACGCT  420

Query  700   ATATTTAACTTCATCACCACCGTGTACGACCTAGACCGGACGCACTCCTTCAAGATGAGT  759
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421   ATATTTAACTTCATCACCACCGTGTACGACCTAGACCGGACGCACTCCTTCAAGATGAGT  480

Query  760   ACGATCGTTCAGGTNCGCAACGTCGATAGCCGGCCTCCGATCTTTAGCCGACCGTTCNCC  819
             |||||||||||||| |||||||||||||||||||||||||||||||||||||||||| ||
Sbjct  481   ACGATCGTTCAGGTACGCAACGTCGATAGCCGGCCTCCGATCTTTAGCCGACCGTTCACC  540

Query  820   AGCGAACGNATCATGNAAANGGAANCATTTTACGCGANCGTGATCGCANTCGACCGTGAC  879
             |||||||| |||||| ||| |||| |||||||||||| |||||||||| ||||||| |||
Sbjct  541   AGCGAACGAATCATGGAAAAGGAACCATTTTACGCGACCGTGATCGCAATCGACCGAGAC  600

Query  880   ACTGGACTAAACAAACCGATCTGTTACGAGCTGACGGCTCTAGTACCGGAATAT---CA-  935
             |||||||||||||||||||||||||||||||||||||||||||| ||||||| |   || 
Sbjct  601   ACTGGACTAAACAAACCGATCTGTTACGAGCTGACGGCTCTAGTCCCGGAATGTAAGCAA  660

Query  936   G-AAA-TATTTCGATATTGGACAAACTGATGGAAAGCTGACCGTGCACCCGATTGATCGA  993
             | ||| |||||||| |||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661   GCAAAATATTTCGAAATTGGACAAACTGATGGAAAGCTGACCGTGCACCCGATTGATCGA  720

Query  994   GATGCCG  1000
             |||||||
Sbjct  721   GATGCCG  727
```

Mismatch

Gap

727 nucleotides retrieved

Entering the Protein Sequence



BLASTP programs search protein databases using a protein query. more...

Enter accession number, gi, or FASTA sequence

```
MKVESWLHLGWLLGLLLVLLP[illegible]VRCQGWGEPRFETGNVENISLAAYNEAQLQQDVVMVEEMDAPFVLL
YINYQ[illegible]SFERTIPESP[illegible]LLDARLQLSEAGRWSIVINRRQDYEVHQRSSLILLAVESTAIPYAIVVNLV
NVLDNAPVMTAQGSCEIEELRGDFVTDCLFNVYHADGFEENGIGNSSTNELSFEIGDVAGARDHFTYV
PSTVTPSQPIYNKLFNLKVLKQLDYTENAIFNFITTVYDLDRTHSFKMSTIVQVRNVDSRPPIFSRPF
XSERIMXXEXFYAXVIAXDRDTGLNKPICYELTALVPEYQKYFDIGQTDGKLTVHPIDRDA
```

Query subrange

From

To

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be sugges □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query (Optional)

Enter an Entrez query to limit search

Algorithm

● blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

□ Show results in a new window

Paste protein sequence

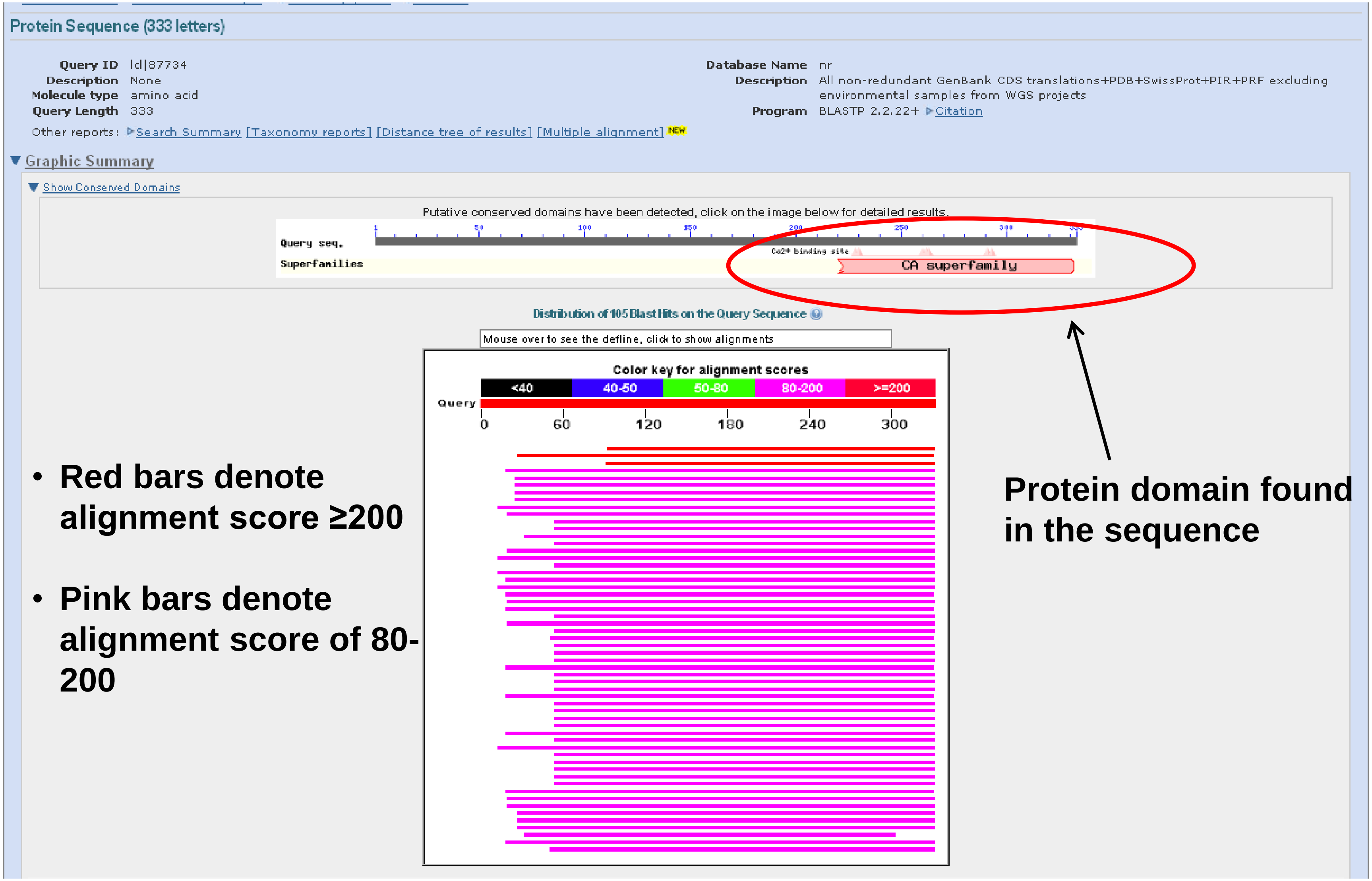
Sequence Identification Numbers

- There are two types of sequence identification numbers, GI and VERSION, each of which have different formats and were implemented at different times.
- GI number (sometimes written in lower case, "gi") is simply a series of digits that are assigned consecutively to each sequence record processed by NCBI.
- The GI number bears no resemblance to the Accession number of the sequence record.
- Nucleotide sequence GI number is shown in the VERSION field of the database record.
- Protein sequence GI number is shown in the CDS/db_xref field of a nucleotide database record, and the VERSION field of a protein database record.

Sequence Identification Numbers

- VERSION is made of the accession number of the database record followed by a dot and a version number (and is, therefore, sometimes referred to as the "accession.version").
- Nucleotide sequence version contains two letters followed by six digits, a dot, and a version number (or for older nucleotide sequence records, the format is one letter followed by five digits, a dot, and a version number)
- Protein sequence version contains three letters followed by five digits, a dot, and a version number.
- The GI number has been used for many years by NCBI to track sequence histories in GenBank and the other sequence databases it maintains.
- The VERSION system of identifiers was adopted in February, 1999, by the International Nucleotide Sequence Database Collaboration (GenBank, the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences; the European Molecular Biology Laboratory, EMBL; the DNA Databank of Japan, DDBJ).
- The two systems of identifiers run in parallel to each other, i.e., when any change is made to a sequence, it receives a new GI number AND an increase to its version number.

Results Obtained



Protein Sequence (333 letters)

Query ID lcl|87734
Description None
Molecule type amino acid
Query Length 333

Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.2.22+ ▸Citation

Other reports: ▸Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] NEW

▼ Graphic Summary

▼ Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.

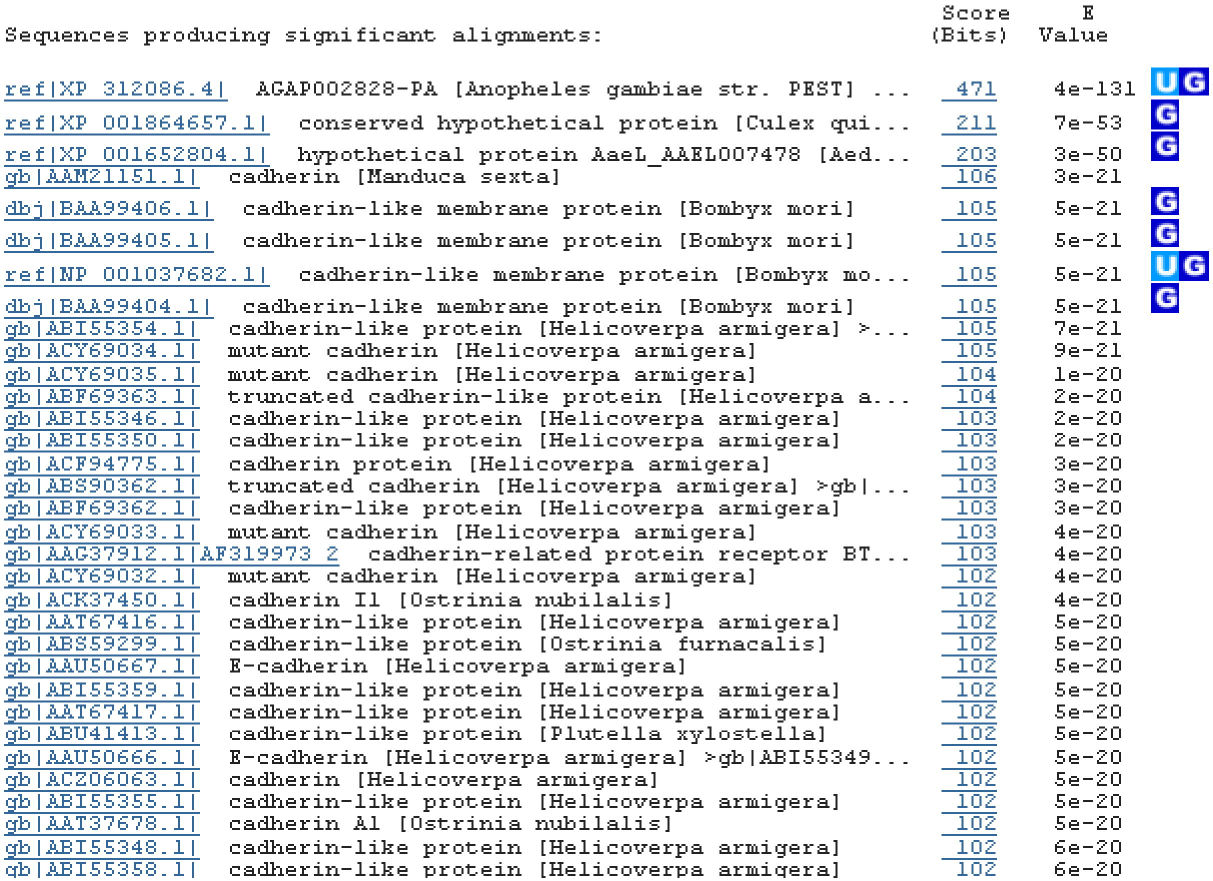
Distribution of 105 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

- **Red bars denote alignment score ≥200**
- **Pink bars denote alignment score of 80-200**

Protein domain found in the sequence

Results Obtained – Proteins in the Database that Matched the Query Sequence



```
                                                                          Score    E
Sequences producing significant alignments:                              (Bits)  Value

ref|XP_312086.4|  AGAP002828-PA [Anopheles gambiae str. PEST] ...          471   4e-131  U G
ref|XP_001864657.1|  conserved hypothetical protein [Culex qui...          211   7e-53   G
ref|XP_001652804.1|  hypothetical protein AaeL_AAEL007478 [Aed...          203   3e-50   G
gb|AAM21151.1|  cadherin [Manduca sexta]                                   106   3e-21
dbj|BAA99406.1|  cadherin-like membrane protein [Bombyx mori]              105   5e-21   G
dbj|BAA99405.1|  cadherin-like membrane protein [Bombyx mori]              105   5e-21   G
ref|NP_001037682.1|  cadherin-like membrane protein [Bombyx mo...          105   5e-21   U G
dbj|BAA99404.1|  cadherin-like membrane protein [Bombyx mori]              105   5e-21   G
gb|ABI55354.1|  cadherin-like protein [Helicoverpa armigera] >...          105   7e-21
gb|ACY69034.1|  mutant cadherin [Helicoverpa armigera]                     105   9e-21
gb|ACY69035.1|  mutant cadherin [Helicoverpa armigera]                     104   1e-20
gb|ABF69363.1|  truncated cadherin-like protein [Helicoverpa a...          104   2e-20
gb|ABI55346.1|  cadherin-like protein [Helicoverpa armigera]               103   2e-20
gb|ABI55350.1|  cadherin-like protein [Helicoverpa armigera]               103   2e-20
gb|ACF94775.1|  cadherin protein [Helicoverpa armigera]                    103   3e-20
gb|ABS90362.1|  truncated cadherin [Helicoverpa armigera] >gb|...          103   3e-20
gb|ABF69362.1|  cadherin-like protein [Helicoverpa armigera]               103   3e-20
gb|ACY69033.1|  mutant cadherin [Helicoverpa armigera]                     103   4e-20
gb|AAG37912.1|AF319973_2  cadherin-related protein receptor BT...          103   4e-20
gb|ACY69032.1|  mutant cadherin [Helicoverpa armigera]                     102   4e-20
gb|ACK37450.1|  cadherin I1 [Ostrinia nubilalis]                           102   4e-20
gb|AAT67416.1|  cadherin-like protein [Helicoverpa armigera]               102   5e-20
gb|ABS59299.1|  cadherin-like protein [Ostrinia furnacalis]                102   5e-20
gb|AAU50667.1|  E-cadherin [Helicoverpa armigera]                          102   5e-20
gb|ABI55359.1|  cadherin-like protein [Helicoverpa armigera]               102   5e-20
gb|AAT67417.1|  cadherin-like protein [Helicoverpa armigera]               102   5e-20
gb|ABU41413.1|  cadherin-like protein [Plutella xylostella]                102   5e-20
gb|AAU50666.1|  E-cadherin [Helicoverpa armigera] >gb|ABI55349...          102   5e-20
gb|ACZ06063.1|  cadherin [Helicoverpa armigera]                            102   5e-20
gb|ABI55355.1|  cadherin-like protein [Helicoverpa armigera]               102   5e-20
gb|AAT37678.1|  cadherin A1 [Ostrinia nubilalis]                           102   5e-20
gb|ABI55348.1|  cadherin-like protein [Helicoverpa armigera]               102   6e-20
gb|ABI55358.1|  cadherin-like protein [Helicoverpa armigera]               102   6e-20
```

Unigene (U) and Entrez Gene (G) links available.

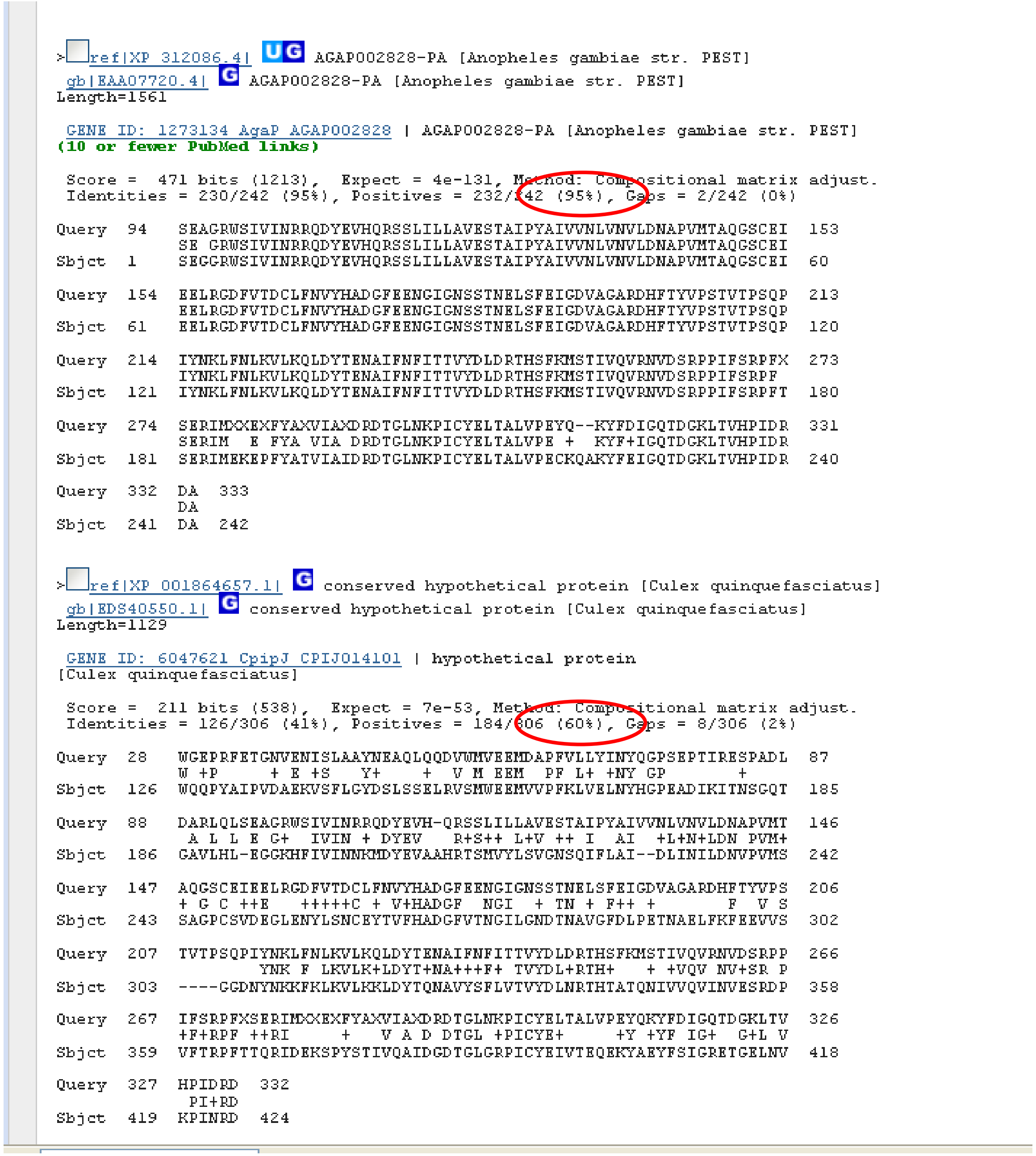
UniGene

- ***UniGene* provides an organized view of the transcriptome.**
- **Each *UniGene* entry is a set of transcript sequences that appear to come from the same transcription locus (gene or expressed pseudogene), together with information on protein similarities, gene expression, cDNA clone reagents, and genomic location.**
- **In addition to sequences of well-characterized genes, hundreds of thousands of novel expressed sequence tag (EST) sequences are available.**
- ***UniGene* is a resource for gene discovery.**
- ***UniGene* has also been used by experimentalists to select reagents for gene mapping projects and large-scale expression analysis.**

Entrez Gene

- Entrez Gene (http://www.ncbi.nim.nih.gov/gene) is the NCBI’s database for gene-specific information.
- Entrez Gene maintains records from genomes which have been completely sequenced, which have an active research community to submit gene-specific information, or which are scheduled for intense sequence analysis.
- The content represents the integration of curation and automated processing from NCBI’s Reference Sequence project (RefSeq), collaborating model organism databases, consortia such as Gene Ontology and other databases within NCBI.
- Records in Entrez Gene are assigned unique, stable and tracked integers as identifiers.
- The content (nomenclature, genomic location, gene products and their attributes, markers, phenotypes and links to citations, sequences, variation details, maps, expression, homologs, protein domains and external databases) is available via interactive browsing through NCBI’s Entrez system, via NCBI’s Entrez programming utilities (E-Utilities) and for bulk transfer by File Transfer Protocol (FTP).

Sequence Comparisons



```
> ref|XP_312086.4|  UG  AGAP002828-PA [Anopheles gambiae str. PEST]
 gb|EAA07720.4|  G  AGAP002828-PA [Anopheles gambiae str. PEST]
Length=1561

 GENE ID: 1273134 AgaP_AGAP002828 | AGAP002828-PA [Anopheles gambiae str. PEST]
(10 or fewer PubMed links)

 Score =  471 bits (1213),  Expect = 4e-131, Method: Compositional matrix adjust.
 Identities = 230/242 (95%), Positives = 232/242 (95%), Gaps = 2/242 (0%)

Query  94   SEAGRWSIVINRRQDYEVHQRSSLILLAVESTAIPYAIVVNLVNVLDNAPVMTAQGSCEI  153
            SE GRWSIVINRRQDYEVHQRSSLILLAVESTAIPYAIVVNLVNVLDNAPVMTAQGSCEI
Sbjct  1    SEGGRWSIVINRRQDYEVHQRSSLILLAVESTAIPYAIVVNLVNVLDNAPVMTAQGSCEI  60

Query  154  EELRGDFVTDCLFNVYHADGFEENGIGNSSTNELSFEIGDVAGARDHFTYVPSTVTPSQP  213
            EELRGDFVTDCLFNVYHADGFEENGIGNSSTNELSFEIGDVAGARDHFTYVPSTVTPSQP
Sbjct  61   EELRGDFVTDCLFNVYHADGFEENGIGNSSTNELSFEIGDVAGARDHFTYVPSTVTPSQP  120

Query  214  IYNKLFNLKVLKQLDYTENAIFNFITTVYDLDRTHSFKMSTIVQVRNVDSRPPIFSRPFX  273
            IYNKLFNLKVLKQLDYTENAIFNFITTVYDLDRTHSFKMSTIVQVRNVDSRPPIFSRPF
Sbjct  121  IYNKLFNLKVLKQLDYTENAIFNFITTVYDLDRTHSFKMSTIVQVRNVDSRPPIFSRPFT  180

Query  274  SERIMXXEXFYAXVIAXDRDTGLNKPICYELTALVPEYQ--KYFDIGQTDGKLTVHPIDR  331
            SERIM  E FYA VIA DRDTGLNKPICYELTALVPE +  KYF+IGQTDGKLTVHPIDR
Sbjct  181  SERIMEKEPFYATVIAIDRDTGLNKPICYELTALVPECKQAKYFEIGQTDGKLTVHPIDR  240

Query  332  DA  333
            DA
Sbjct  241  DA  242


> ref|XP_001864657.1|  G  conserved hypothetical protein [Culex quinquefasciatus]
 gb|EDS40550.1|  G  conserved hypothetical protein [Culex quinquefasciatus]
Length=1129

 GENE ID: 6047621 CpipJ_CPIJ014101 | hypothetical protein
[Culex quinquefasciatus]

 Score =  211 bits (538),  Expect = 7e-53, Method: Compositional matrix adjust.
 Identities = 126/306 (41%), Positives = 184/306 (60%), Gaps = 8/306 (2%)

Query  28   WGEPRFETGNVENISLAAYNEAQLQQDVWMVEEMDAPFVLLYINYQGPSEPTIRESPADL  87
            W +P     + E +S   Y+    +  V M EEM  PF L+ +NY GP       +
Sbjct  126  WQQPYAIPVDAEKVSFLGYDSLSSELRVSMWEEMVVPFKLVELNYHGPEADIKITNSGQT  185

Query  88   DARLQLSEAGRWSIVINRRQDYEVH-QRSSLILLAVESTAIPYAIVVNLVNVLDNAPVMT  146
             A L L E G+  IVIN + DYEV   R+S++ L+V ++ I  AI  +L+N+LDN PVM+
Sbjct  186  GAVLHL-EGGKHFIVINNKMDYEVAAHRTSMVYLSVGNSQIFLAI--DLINILDNVPVMS  242

Query  147  AQGSCEIEELRGDFVTDCLFNVYHADGFEENGIGNSSTNELSFEIGDVAGARDHFTYVPS  206
            + G C ++E   +++++C + V+HADGF  NGI  + TN + F++ +       F  V S
Sbjct  243  SAGPCSVDEGLENYLSNCEYTVFHADGFVTNGILGNDTNAVGFDLPETNAELFKFEEVVS  302

Query  207  TVTPSQPIYNKLFNLKVLKQLDYTENAIFNFITTVYDLDRTHSFKMSTIVQVRNVDSRPP  266
                    YNK F LKVLK+LDYT+NA+++F+ TVYDL+RTH+   + +VQV NV+SR P
Sbjct  303  ----GGDNYNKKFKLKVLKKLDYTQNAVYSFLVTVYDLNRTHTATQNIVVQVINVESRDP  358

Query  267  IFSRPFXSERIMXXEXFYAXVIAXDRDTGLNKPICYELTALVPEYQKYFDIGQTDGKLTV  326
            +F+RPF ++RI     +   V A D DTGL +PICYE+     +Y +YF IG+  G+L V
Sbjct  359  VFTRPFTTQRIDEKSPYSTIVQAIDGDTGLGRPICYEIVTEQEKYAEYFSIGRETGELNV  418

Query  327  HPIDRD  332
             PI+RD
Sbjct  419  KPINRD  424
```

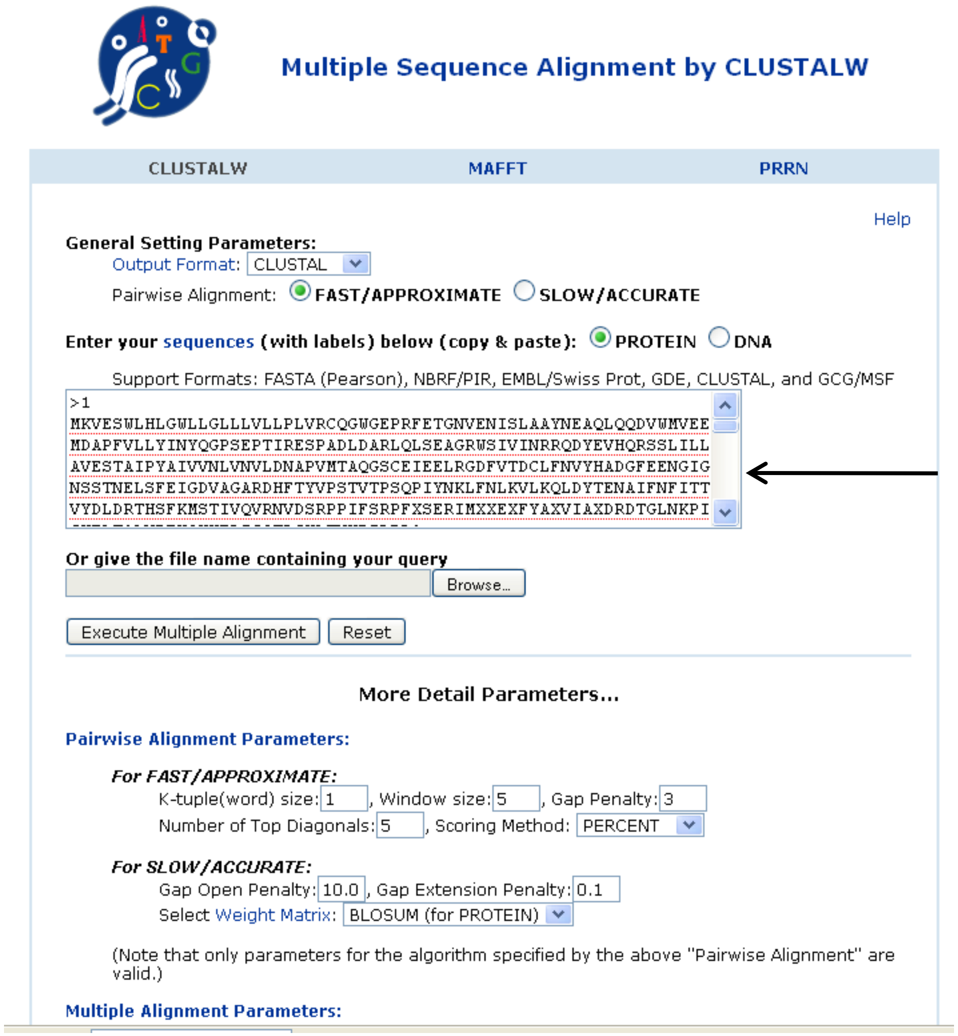
CLUSTAL

- **CLUSTAL is a widely used multiple sequence alignment computer program.**
- **There are two main variations:**
  - ✓**ClustalW: command line interface**
  - ✓**ClustalX: has a graphical user interface**

CLUSTALW

- The basic information that multiple alignments of protein sequences provide is identification of conserved sequence regions.
- This is very useful in designing experiments to test and modify the function of specific proteins, in predicting the function and structure of proteins and in identifying new members of protein families.
- In ClustalW, a pairwise score is calculated for every pair of sequences that are to be aligned.
- Pairwise scores are calculated as the number of identities in the best alignment divided by the number of residues compared (gap positions are excluded).
- As the pairwise score is calculated independently of the matrix and gaps chosen, it will always be the same value for a particular pair of sequences.
- Alignment score is calculated in two ways – fast and slow (more accurate mode).
- The scores are calculated from separate pairwise alignments.
- The scores can be calculated using two methods: dynamic programming (slow but accurate) or by the method of Wilbur and Lipman (extremely fast but approximate).
- The Wilbur-Lipman Method constructs tables of prime K-tuples to find regions of similarity between two or more DNA sequence pairs.
- Prime *k*-tuple is a finite collection of values representing a repeatable pattern of differences between prime numbers.

CLUSTALW



Multiple Sequence Alignment by CLUSTALW

General Setting Parameters:

Output Format: CLUSTAL

Pairwise Alignment: ● FAST/APPROXIMATE ○ SLOW/ACCURATE

Enter your sequences (with labels) below (copy & paste): ● PROTEIN ○ DNA

Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF

```
>1
MKVESWLHLGWLLGLLLVLLPLVRCQGWGEPRFETGNVENISLAAYNEAQLQQDVWMVEE
MDAPFVLLYINYQGPSEPTIRESPADLDARLQLSEAGRWSIVINRRQDYEVHQRSSLILL
AVESTAIPYAIVVNLVNVLDNAPVMTAQGSCEIEELRGDFVTDCLFNVYHADGFEENGIG
NSSTNELSFEIGDVAGARDHFTYVPSTVTPSQPIYNKLFNLKVLKQLDYTENAIFNFITT
VYDLDRTHSFKMSTIVQVRNVDSRPPIFSRPFXSERIMXXEXFYAXVIAXDRDTGLNKPI
```

Or give the file name containing your query

More Detail Parameters...

Pairwise Alignment Parameters:

For FAST/APPROXIMATE:

K-tuple(word) size: 1, Window size: 5, Gap Penalty: 3

Number of Top Diagonals: 5, Scoring Method: PERCENT

For SLOW/ACCURATE:

Gap Open Penalty: 10.0, Gap Extension Penalty: 0.1

Select Weight Matrix: BLOSUM (for PROTEIN)

(Note that only parameters for the algorithm specified by the above "Pairwise Alignment" are valid.)

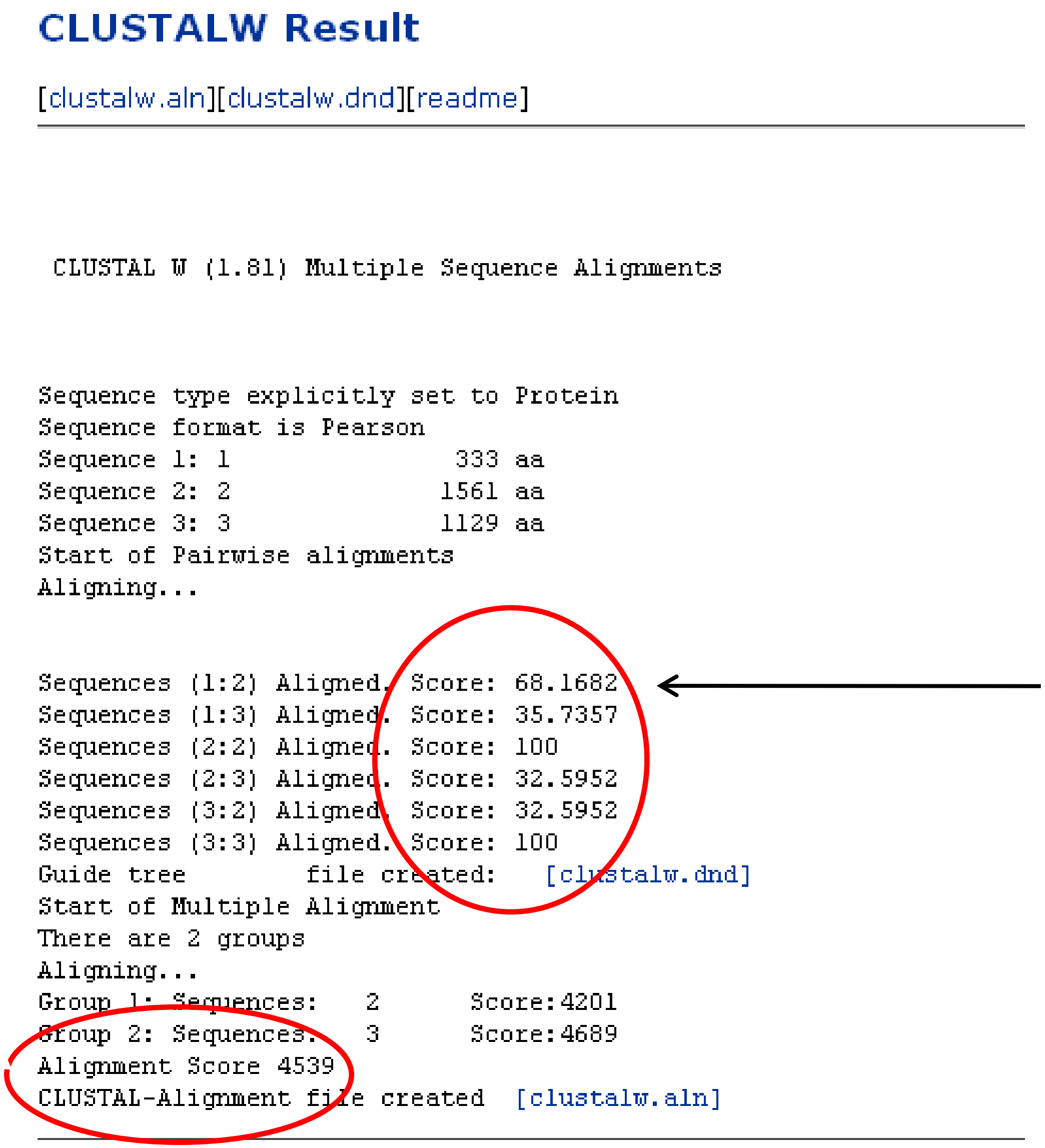
Multiple Alignment Parameters:

Paste all sequences in FASTA format. Three sequences were used:

1. **Anopheles (derived)**
2. **Anopheles (from database)**
3. **Culex (from database).**

FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes.

Results Obtained



CLUSTALW Result

[clustalw.aln][clustalw.dnd][readme]

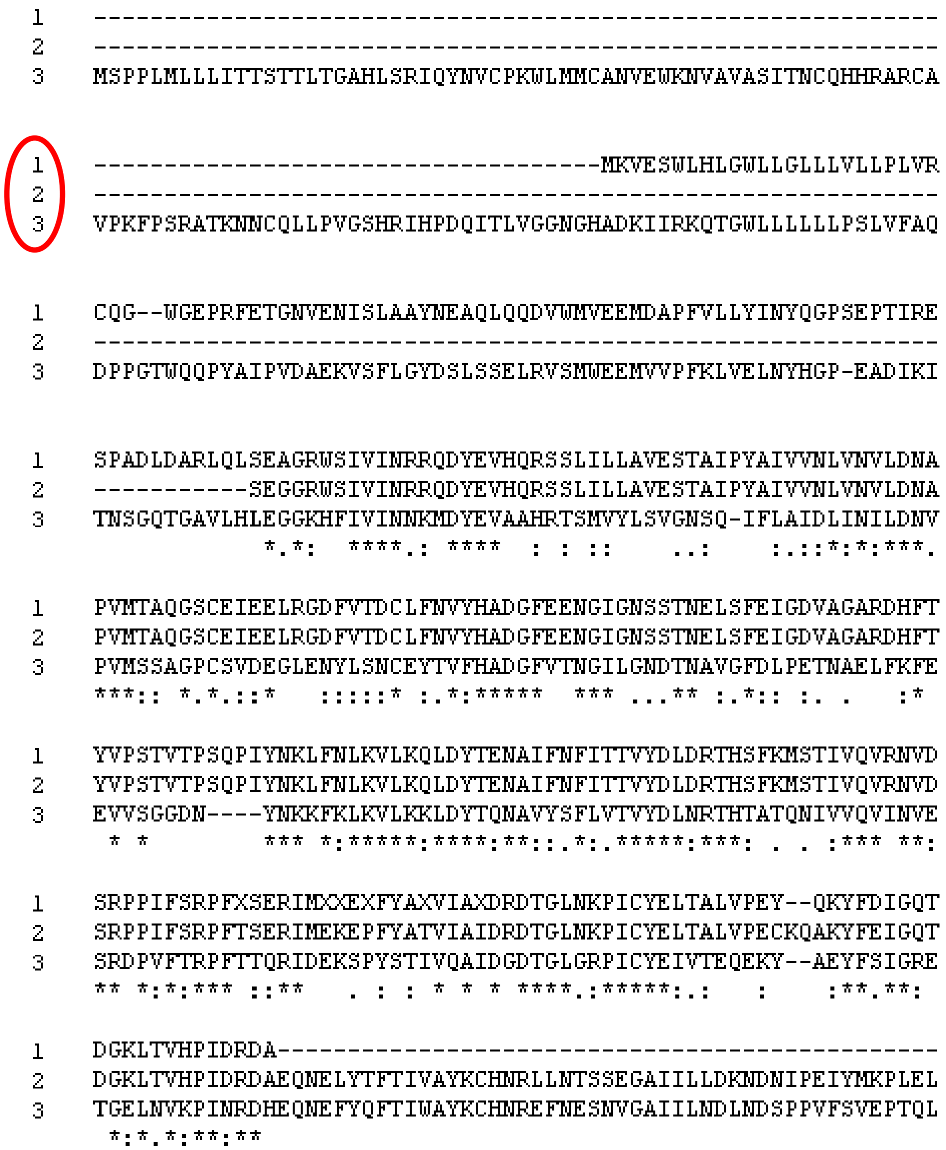
```
 CLUSTAL W (1.81) Multiple Sequence Alignments



Sequence type explicitly set to Protein
Sequence format is Pearson
Sequence 1: 1              333 aa
Sequence 2: 2             1561 aa
Sequence 3: 3             1129 aa
Start of Pairwise alignments
Aligning...

Sequences (1:2) Aligned. Score: 68.1682
Sequences (1:3) Aligned. Score: 35.7357
Sequences (2:2) Aligned. Score: 100
Sequences (2:3) Aligned. Score: 32.5952
Sequences (3:2) Aligned. Score: 32.5952
Sequences (3:3) Aligned. Score: 100
Guide tree        file created:   [clustalw.dnd]
Start of Multiple Alignment
There are 2 groups
Aligning...
Group 1: Sequences:   2      Score:4201
Group 2: Sequences:   3      Score:4689
Alignment Score 4539
CLUSTAL-Alignment file created  [clustalw.aln]
```

Pairwise scores calculated for each pair of sequences

Alignment score calculated from pairwise scores

Sequence Comparisons

Three sequences compared



```
1     ------------------------------------------------------------
2     ------------------------------------------------------------
3     MSPPLMLLLITTSTTLTGAHLSRIQYNVCPKWLMMCANVEWKNVAVASITNCQHHRARCA


1     ------------------------------------MKVESWLHLGWLLGLLLVLLPLVR
2     ------------------------------------------------------------
3     VPKFPSRATKNNCQLLPVGSHRIHPDQITLVGGNGHADKIIRKQTGWLLLLLLPSLVFAQ


1     CQG--WGEPRFETGNVENISLAAYNEAQLQQDVWMVEEMDAPFVLLYINYQGPSEPTIRE
2     ------------------------------------------------------------
3     DPPGTWQQPYAIPVDAEKVSFLGYDSLSSELRVSMWEEMVVPFKLVELNYHGP-EADIKI


1     SPADLDARLQLSEAGRWSIVINRRQDYEVHQRSSLILLAVESTAIPYAIVVNLVNVLDNA
2     -----------SEGGRWSIVINRRQDYEVHQRSSLILLAVESTAIPYAIVVNLVNVLDNA
3     TNSGQTGAVLHLEGGKHFIVINNKMDYEVAAHRTSMVYLSVGNSQ-IFLAIDLINILDNV
                  *.*:  ****.: ****  : : ::    ..:    :.::*:*:***.

1     PVMTAQGSCEIEELRGDFVTDCLFNVYHADGFEENGIGNSSTNELSFEIGDVAGARDHFT
2     PVMTAQGSCEIEELRGDFVTDCLFNVYHADGFEENGIGNSSTNELSFEIGDVAGARDHFT
3     PVMSSAGPCSVDEGLENYLSNCEYTVFHADGFVTNGILGNDTNAVGFDLPETNAELFKFE
      ***:: *.*.::*   :::::* :.*:*****  *** ...** :.*:: :. .   :*

1     YVPSTVTPSQPIYNKLFNLKVLKQLDYTENAIFNFITTVYDLDRTHSFKMSTIVQVRNVD
2     YVPSTVTPSQPIYNKLFNLKVLKQLDYTENAIFNFITTVYDLDRTHSFKMSTIVQVRNVD
3     EVVSGGDN----YNKKFKLKVLKKLDYTQNAVYSFLVTVYDLNRTHTATQNIVVQVINVE
       * *        *** *:*****:****:**::.*:.*****:***: . . :*** **:

1     SRPPIFSRPFXSERIMXXEXFYAXVIAXDRDTGLNKPICYELTALVPEY--QKYFDIGQT
2     SRPPIFSRPFTSERIMEKEPFYATVIAIDRDTGLNKPICYELTALVPECKQAKYFEIGQT
3     SRDPVFTRPFTTQRIDEKSPYSTIVQAIDGDTGLGRPICYEIVTEQEKY--AEYFSIGRE
      ** *:*:*** ::**   . : : * * * ****.:*****:.:   :    :**.**:

1     DGKLTVHPIDRDA-----------------------------------------------
2     DGKLTVHPIDRDAEQNELYTFTIVAYKCHNRLLNTSSEGAIILLDKNDNIPEIYMKPLEL
3     TGELNVKPINRDHEQNEFYQFTIWAYKCHNREFNESNVGAIILNDLNDSPPVFSVEPTQL
       *:*.*:**:**
```

Symbols Used in CLUSTALW:

- ✓ *** Indicates identical amino acid residues in all sequences (or identical bases if DNA sequences are aligned)**
- ✓ **: indicates different but highly conserved (very similar) amino acids**
- ✓ **. Indicates different amino acids that are somewhat similar**
- ✓ **Blank indicates dissimilar amino acids or gaps (or different bases if DNA sequences are aligned)**

Dayhoff Point Accepted Mutation (PAM)

- Point Accepted Mutation (PAM) is a set of matrices used to score sequence alignments.
- The PAM matrices are based on 1572 observed mutations in 71 families of closely related proteins.
- Each matrix is twenty-by-twenty (for the twenty standard amino acids).
- The value in a given cell represents the probability of a substitution of one amino acid for another.
- This type of matrix is commonly known as a substitution matrix.
- The PAM matrices imply a Markov chain model of protein mutation.
- A Markov model is a stochastic model used to model randomly changing systems where it is assumed that future states depend only on the current state, not on the events that occurred before it
- The PAM matrices are normalized so that the PAM1 matrix has one point mutation per hundred amino acids, and is appropriate for scoring sequences which are very similar.
- PAM matrices for comparing sequences of lower similarity are calculated from repeated multiplication of the PAM1 matrix by itself.
- PAM250 is equivalent to 250 substitutions per hundred amino acids.

BLOck SUbstitution Matrix (BLOSUM)

- PAM compares closely related species and does not work very well for aligning evolutionarily divergent sequences.
- Sequence changes over long evolutionary time-scales are not approximated very efficiently by compounding small changes that occur over short time-scales.
- The BLOSUM series of matrices uses multiple alignments of evolutionarily divergent proteins and the probabilities used in the matrix calculation are computed by looking at "blocks" of conserved sequences found in multiple protein alignments.
- These conserved sequences are assumed to be of functional importance within related proteins.
- To reduce bias from closely related sequences, segments in a block with a sequence identity above a certain threshold are clustered giving weight value of 1 to each such cluster.
- For the BLOSUM45 matrix, this threshold is set at 45%.
- Pairs frequencies are counted between clusters and, therefore, pairs are counted only between segments less than 45% identical.
- A higher numbered BLOSUM matrix is used for aligning two closely related sequences and a lower number for more divergent sequences.
- BLOSUM62 matrix does an excellent job detecting similarities in distant sequences, and is the default matrix used most for alignment applications such as BLAST.

Differences Between PAM and BLOSUM

- PAM matrices are based on an explicit evolutionary model (i.e., replacements are counted on the branches of a phylogenetic tree) whereas the BLOSUM matrices are based on an implicit model of evolution.
- The PAM matrices are based on mutations observed throughout a global alignment and includes both highly conserved and highly mutable regions.
- The BLOSUM matrices are based only on highly conserved regions in series of alignments forbidden to contain gaps.
- The method used to count the replacements is different: unlike the PAM matrix, the BLOSUM procedure uses groups of sequences within which not all mutations are counted the same.
- Higher numbers in the PAM matrix-naming scheme denote larger evolutionary distance whereas higher numbers in the BLOSUM matrix-naming scheme denote higher sequence similarity and, therefore, smaller evolutionary distance.
- Example: PAM150 is used for more distant sequences than PAM100; BLOSUM62 is used for closer sequences than Blosum50.