BIOL 419/519 Bioinformatics Research - June 4, 2013
Introduction to BLAST
(adapted from Wilson Leung, 02/09/2013)

Please read **Detecting and Interpreting Genetic Homology: Lecture Notes on Alignment.**

Introduction

The Basic Local Alignment Search Tool (BLAST) is a program that can detect sequence similarity between a query sequence and sequences within a database. The ability to detect sequence homology allows us to identify putative genes in a novel sequence. It also allows us to determine if a gene or protein is related to other known genes or proteins.

BLAST is popular because it can quickly identify regions of local similarity between two sequences. More importantly, BLAST uses a robust statistical framework that can determine if the alignment between two sequences is statistically significant. In this tutorial, we will use the BLAST web interface at the National Center for Biotechnology Information (NCBI) to explore a sequence from *Drosophila melanogaster*.

Resources

Course website:

http://www.discoveryandinnovation.com/bioinformatics/

From the main page, click the **Tools** link at the top of the page for access to links to various tools:

http://www.discoveryandinnovation.com/bioinformatics/tools.html

The **Class Projects** link gives access to this document and the required data file:

http://www.discoveryandinnovation.com/bioinformatics/class_projects2.html

The **Glossary** defines many terms in genomics and bioinformatics.

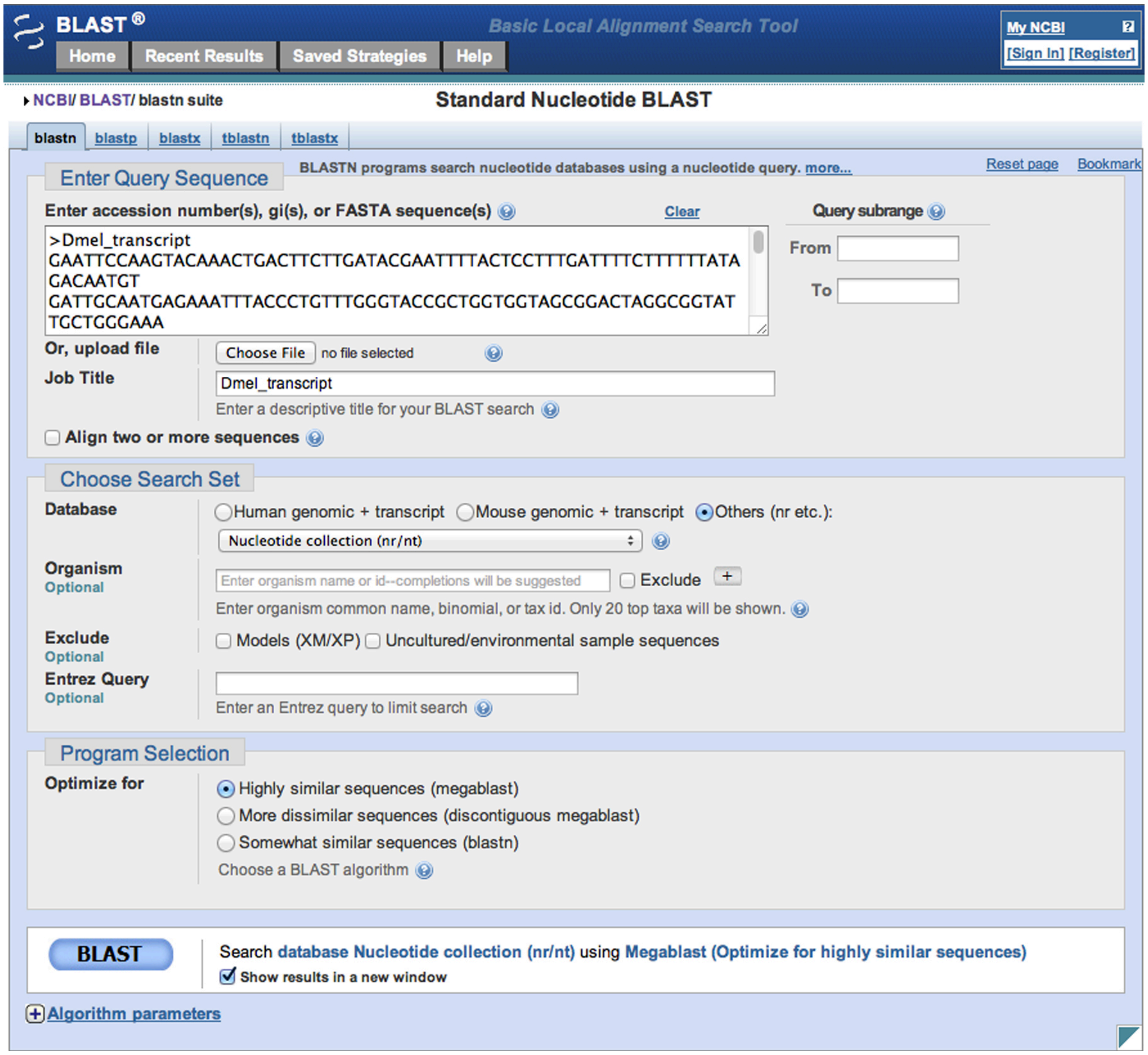
1. Identification of the sequence

From the **Class Projects** page, click the link to the **data file** to recover the sequence of a cDNA from *Drosophila melanogaster*. Copy the sequence by selecting it, then entering Command-c (copy) on your Mac laptop, or the equivalent command on a PC.

From the Tools page, click the link to **blastn** from the first line of BLAST tools. This takes you to NCBI BLAST. Paste the transcript sequence into the query box (Command-v) as shown in the screenshot on the next page.

The screenshot below shows the NCBI blastn page with transcript sequence loaded. The default database on this screen is **Nucleotide collection (nr)**. We have the option to change the **Program Selection** (to allow for more inexact matches) and the **Algorithm parameters** (we will leave these at the default settings now).

Check the box near the BLAST button to **Show results in a new window** and click **BLAST**.



Standard Nucleotide BLAST

Enter Query Sequence

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

Enter accession number(s), gi(s), or FASTA sequence(s)

```
>Dmel_transcript
GAATTCCAAGTACAAACTGACTTCTTGATACGAATTTTACTCCTTTGATTTTCTTTTTTATA
GACAATGT
GATTGCAATGAGAAATTTACCCTGTTTGGGTACCGCTGGTGGTAGCGGACTAGGCGGTAT
TGCTGGGAAA
```

Query subrange

From

To

Or, upload file: no file selected

Job Title: Dmel_transcript

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:

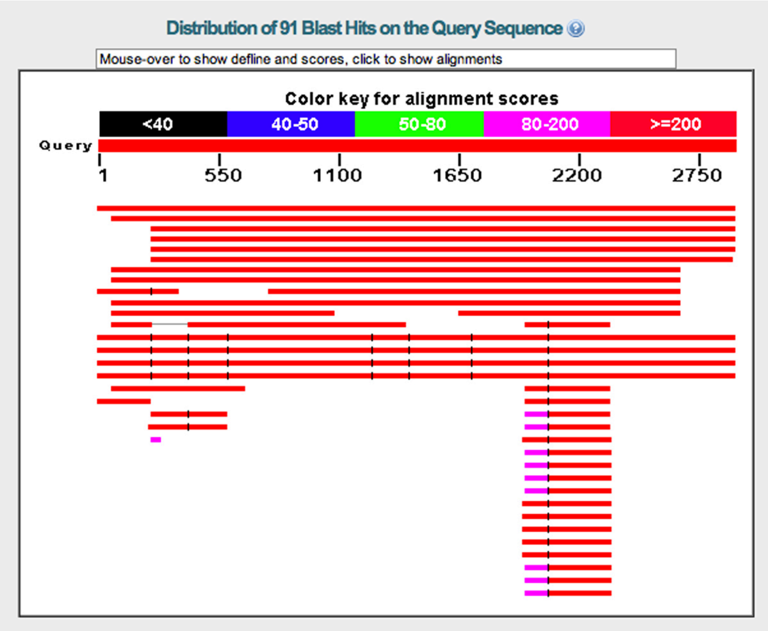
● Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

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

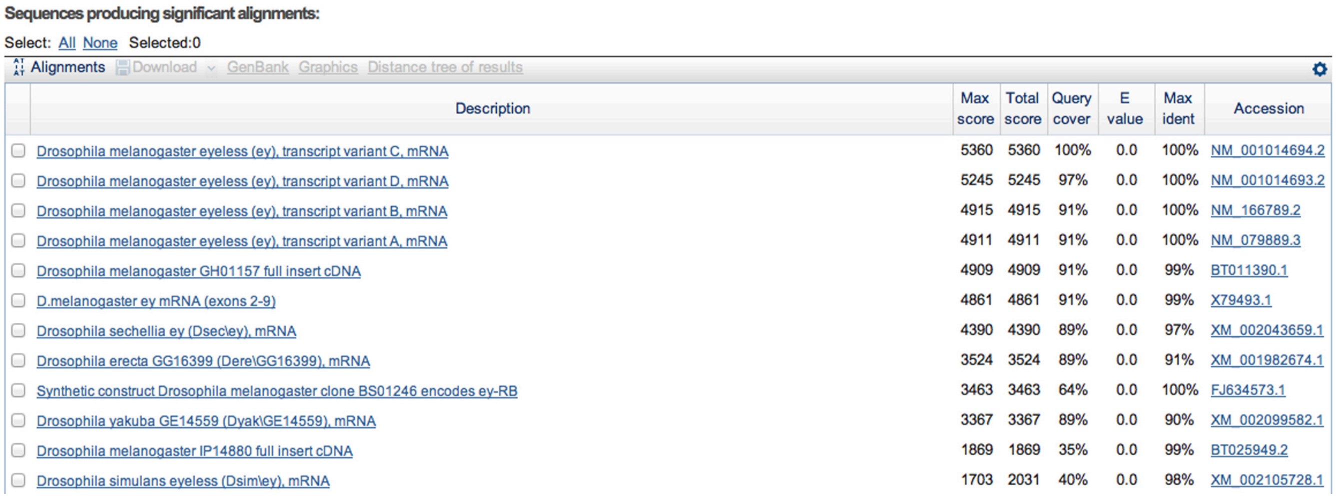
☑ Show results in a new window

The first results are presented visually, as shown below. Each line represents a matching sequence. The sequence at the top of the figure aligns across the entire length of the query sequence. Subsequent matches align only partially, until at the bottom of the figure we see sequences that align for a small portion of the query sequence, with color used to indicate lower-quality alignments (lower scores).


Distribution of 91 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
550
1100
1650
2200
2750

The next part of the results shows the sequence descriptions, scores, E values, and Accession IDs. Click any description line to jump to the alignment, or click the Accession ID to find the entry in GenBank.

Notice that all of the E values are highly significant (less than e-150 or so). You can see that the alignments are almost perfect for the first few matches.



Sequences producing significant alignments:

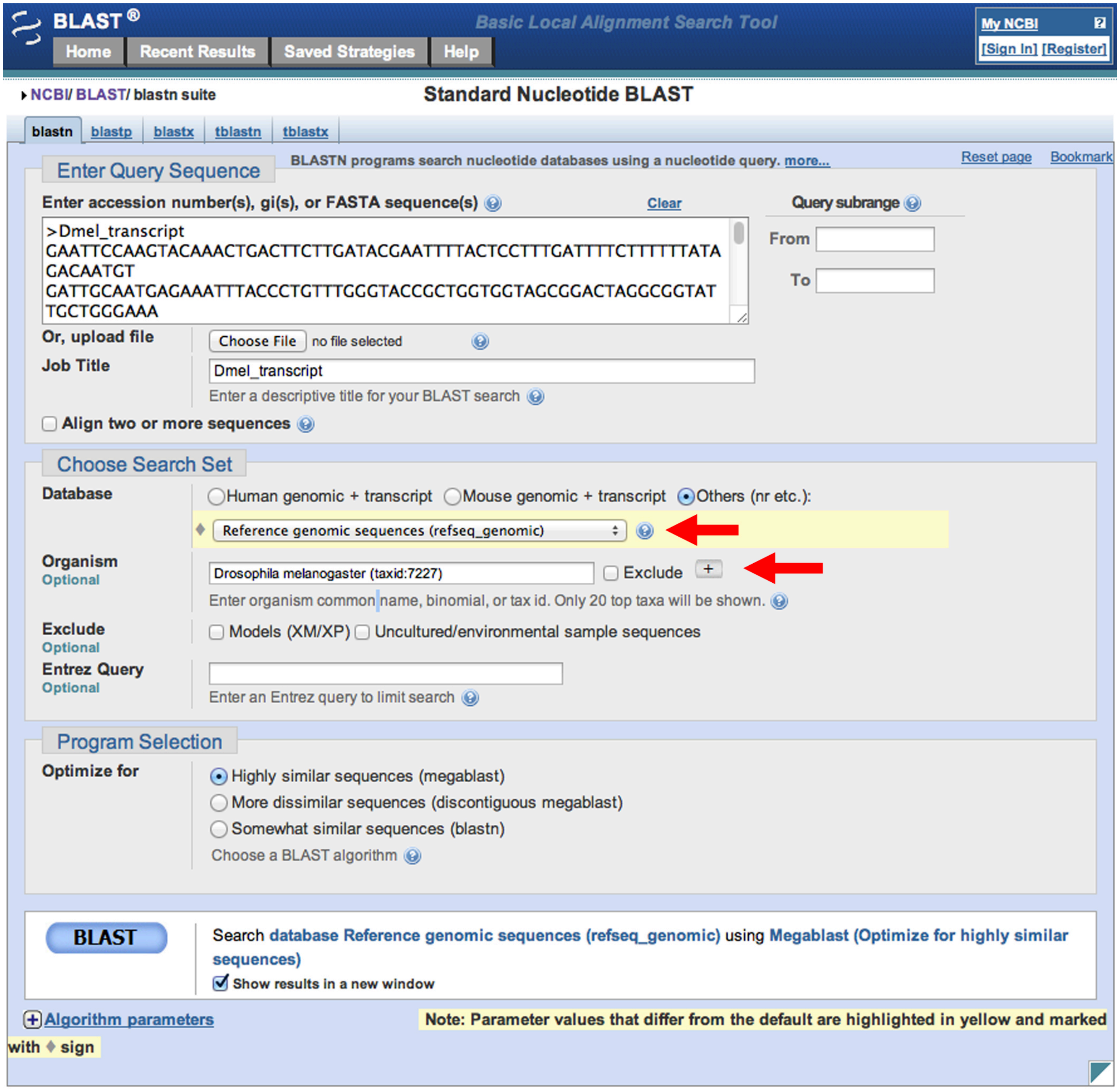
Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
☐	Drosophila melanogaster eyeless (ey), transcript variant C, mRNA	5360	5360	100%	0.0	100%	NM_001014694.2
☐	Drosophila melanogaster eyeless (ey), transcript variant D, mRNA	5245	5245	97%	0.0	100%	NM_001014693.2
☐	Drosophila melanogaster eyeless (ey), transcript variant B, mRNA	4915	4915	91%	0.0	100%	NM_166789.2
☐	Drosophila melanogaster eyeless (ey), transcript variant A, mRNA	4911	4911	91%	0.0	100%	NM_079889.3
☐	Drosophila melanogaster GH01157 full insert cDNA	4909	4909	91%	0.0	99%	BT011390.1
☐	D.melanogaster ey mRNA (exons 2-9)	4861	4861	91%	0.0	99%	X79493.1
☐	Drosophila sechellia ey (Dsec\ey), mRNA	4390	4390	89%	0.0	97%	XM_002043659.1
☐	Drosophila erecta GG16399 (Dere\GG16399), mRNA	3524	3524	89%	0.0	91%	XM_001982674.1
☐	Synthetic construct Drosophila melanogaster clone BS01246 encodes ey-RB	3463	3463	64%	0.0	100%	FJ634573.1
☐	Drosophila yakuba GE14559 (Dyak\GE14559), mRNA	3367	3367	89%	0.0	90%	XM_002099582.1
☐	Drosophila melanogaster IP14880 full insert cDNA	1869	1869	35%	0.0	99%	BT025949.2
☐	Drosophila simulans eyeless (Dsim\ey), mRNA	1703	2031	40%	0.0	98%	XM_002105728.1

2. Mapping the sequence to the genome

Let's use BLAST to map the cDNA sequence to the genome assembly of *Drosophila melanogaster*. Go back to the blastn screen, making the following changes to the settings: 1. Change the database to **Reference genomic sequences (refseq_genomic)**. 2. Enter **Drosophila melanogaster** as the organism (the field will auto-fill to offer you selections once you start typing).



Standard Nucleotide BLAST

Enter Query Sequence

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

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
>Dmel_transcript
GAATTCCAAGTACAAACTGACTTCTTGATACGAATTTTACTCCTTTGATTTTCTTTTTTATA
GACAATGT
GATTGCAATGAGAAATTTACCCTGTTTGGGTACCGCTGGTGGTAGCGGACTAGGCGGTAT
TGCTGGGAAA
```

Query subrange: From ___ To ___

Or, upload file: Choose File no file selected

Job Title: Dmel_transcript

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.):

♦ Reference genomic sequences (refseq_genomic)

Organism (Optional): Drosophila melanogaster (taxid:7227) □ 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

BLAST

Search database Reference genomic sequences (refseq_genomic) using Megablast (Optimize for highly similar sequences)

☑ Show results in a new window

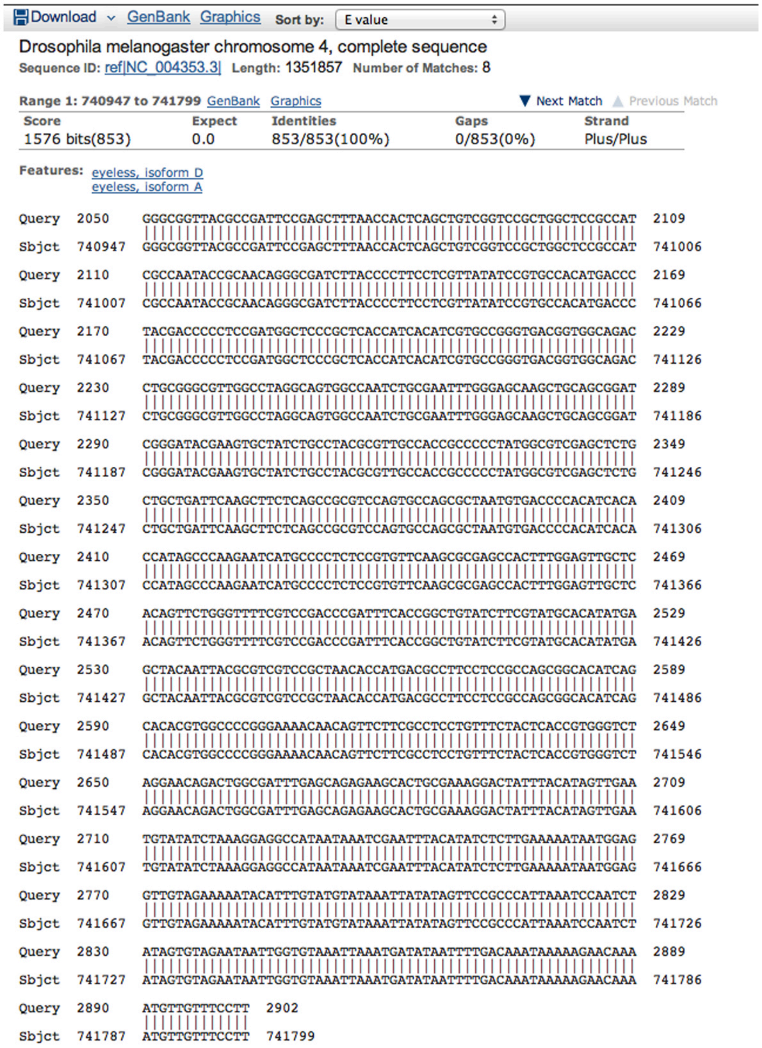
+ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

Click **BLAST**.

Let’s skip straight to the alignments. The alignments are broken into segments, each matches the query sequence perfectly.

In the first alignment, bases 2050-2902 of the query sequence align to 740947-741799 of the subject sequence (the genomic assembly of chromosome 4). Find the part of the genomic sequence that aligns to the first segment of the cDNA sequence. How many aligning segments are there? We will pause as a class to tabulate and discuss the results.



Drosophila melanogaster chromosome 4, complete sequence
Sequence ID: ref|NC_004353.3| Length: 1351857 Number of Matches: 8

Range 1: 740947 to 741799 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
1576 bits(853)	0.0	853/853(100%)	0/853(0%)	Plus/Plus

Features: eyeless, isoform D
eyeless, isoform A

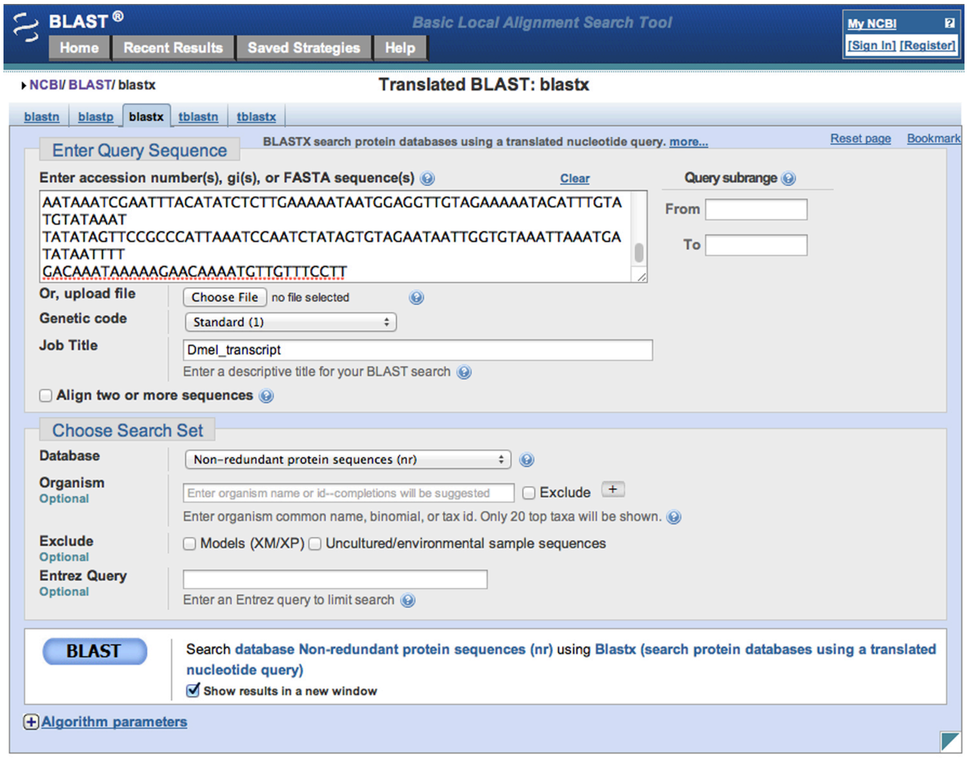
```
Query  2050    GGGCGGTTACGCCGATTCCGAGCTTTAACCACTCAGCTGTCGGTCCGCTGGCTCCGCCAT  2109
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  740947  GGGCGGTTACGCCGATTCCGAGCTTTAACCACTCAGCTGTCGGTCCGCTGGCTCCGCCAT  741006

Query  2110    CGCCAATACCGCAACAGGGCGATCTTACCCCTTCCTCGTTATATCCGTGCCACATGACCC  2169
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  741007  CGCCAATACCGCAACAGGGCGATCTTACCCCTTCCTCGTTATATCCGTGCCACATGACCC  741066

Query  2170    TACGACCCCCTCCGATGGCTCCCGCTCACCATCACATCGTGCCGGGTGACGGTGGCAGAC  2229
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  741067  TACGACCCCCTCCGATGGCTCCCGCTCACCATCACATCGTGCCGGGTGACGGTGGCAGAC  741126

Query  2230    CTGCGGGCGTTGGCCTAGGCAGTGGCCAATCTGCGAATTTGGGAGCAAGCTGCAGCGGAT  2289
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  741127  CTGCGGGCGTTGGCCTAGGCAGTGGCCAATCTGCGAATTTGGGAGCAAGCTGCAGCGGAT  741186

Query  2290    CGGGATACGAAGTGCTATCTGCCTACGCGTTGCCACCGCCCCCTATGGCGTCGAGCTCTG  2349
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  741187  CGGGATACGAAGTGCTATCTGCCTACGCGTTGCCACCGCCCCCTATGGCGTCGAGCTCTG  741246

Query  2350    CTGCTGATTCAAGCTTCTCAGCCGCGTCCAGTGCCAGCGCTAATGTGACCCCACATCACA  2409
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  741247  CTGCTGATTCAAGCTTCTCAGCCGCGTCCAGTGCCAGCGCTAATGTGACCCCACATCACA  741306

Query  2410    CCATAGCCCAAGAATCATGCCCCTCTCCGTGTTCAAGCGCGAGCCACTTTGGAGTTGCTC  2469
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  741307  CCATAGCCCAAGAATCATGCCCCTCTCCGTGTTCAAGCGCGAGCCACTTTGGAGTTGCTC  741366

Query  2470    ACAGTTCTGGGTTTTCGTCCGACCCGATTTCACCGGCTGTATCTTCGTATGCACATATGA  2529
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  741367  ACAGTTCTGGGTTTTCGTCCGACCCGATTTCACCGGCTGTATCTTCGTATGCACATATGA  741426

Query  2530    GCTACAATTACGCGTCGTCCGCTAACACCATGACGCCTTCCTCCGCCAGCGGCACATCAG  2589
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  741427  GCTACAATTACGCGTCGTCCGCTAACACCATGACGCCTTCCTCCGCCAGCGGCACATCAG  741486

Query  2590    CACACGTGGCCCCGGGAAAACAACAGTTCTTCGCCTCCTGTTTCTACTCACCGTGGGTCT  2649
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  741487  CACACGTGGCCCCGGGAAAACAACAGTTCTTCGCCTCCTGTTTCTACTCACCGTGGGTCT  741546

Query  2650    AGGAACAGACTGGCGATTTGAGCAGAGAAGCACTGCGAAAGGACTATTTACATAGTTGAA  2709
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  741547  AGGAACAGACTGGCGATTTGAGCAGAGAAGCACTGCGAAAGGACTATTTACATAGTTGAA  741606

Query  2710    TGTATATCTAAAGGAGGCCATAATAAATCGAATTTACATATCTCTTGAAAAATAATGGAG  2769
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  741607  TGTATATCTAAAGGAGGCCATAATAAATCGAATTTACATATCTCTTGAAAAATAATGGAG  741666

Query  2770    GTTGTAGAAAAATACATTTGTATGTATAAATTATATAGTTCCGCCCATTAAATCCAATCT  2829
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  741667  GTTGTAGAAAAATACATTTGTATGTATAAATTATATAGTTCCGCCCATTAAATCCAATCT  741726

Query  2830    ATAGTGTAGAATAATTGGTGTAAATTAAATGATATAATTTTGACAAATAAAAAGAACAAA  2889
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  741727  ATAGTGTAGAATAATTGGTGTAAATTAAATGATATAATTTTGACAAATAAAAAGAACAAA  741786

Query  2890    ATGTTGTTTCCTT  2902
               |||||||||||||
Sbjct  741787  ATGTTGTTTCCTT  741799
```

3. Obtain the protein sequence corresponding to the transcript

Return to the BLAST page. Use the tabs to select blastx, which will allow us to use a translation of the transcript to search a protein sequence database.

Paste in the cDNA sequence as before.

Check the box to Show results in a new window, and click **BLAST**.



Translated BLAST: blastx

Enter Query Sequence

BLASTX search protein databases using a translated nucleotide query. more...

Enter accession number(s), gi(s), or FASTA sequence(s)

```
AATAAATCGAATTTACATATCTCTTGAAAAATAATGGAGGTTGTAGAAAAATACATTTGTA
TGTATAAAT
TATATAGTTCCGCCCATTAAATCCAATCTATAGTGTAGAATAATTGGTGTAAATTAAATGA
TATAATTTT
GACAAATAAAAAGAACAAAATGTTGTTTCCTT
```

Query subrange

From

To

Or, upload file: no file selected

Genetic code: Standard (1)

Job Title: Dmel_transcript

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

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

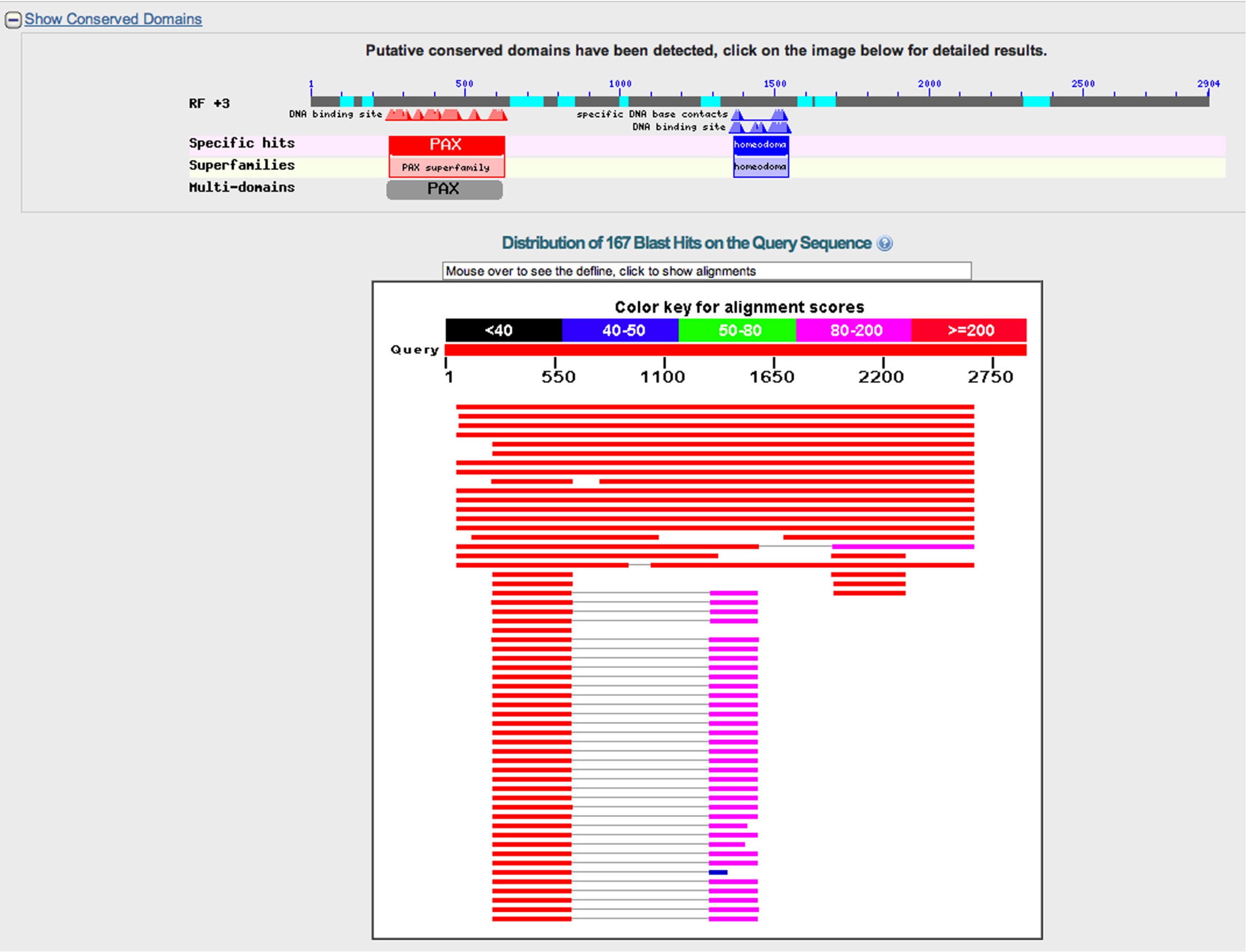
Enter an Entrez query to limit search

Search database Non-redundant protein sequences (nr) using Blastx (search protein databases using a translated nucleotide query)

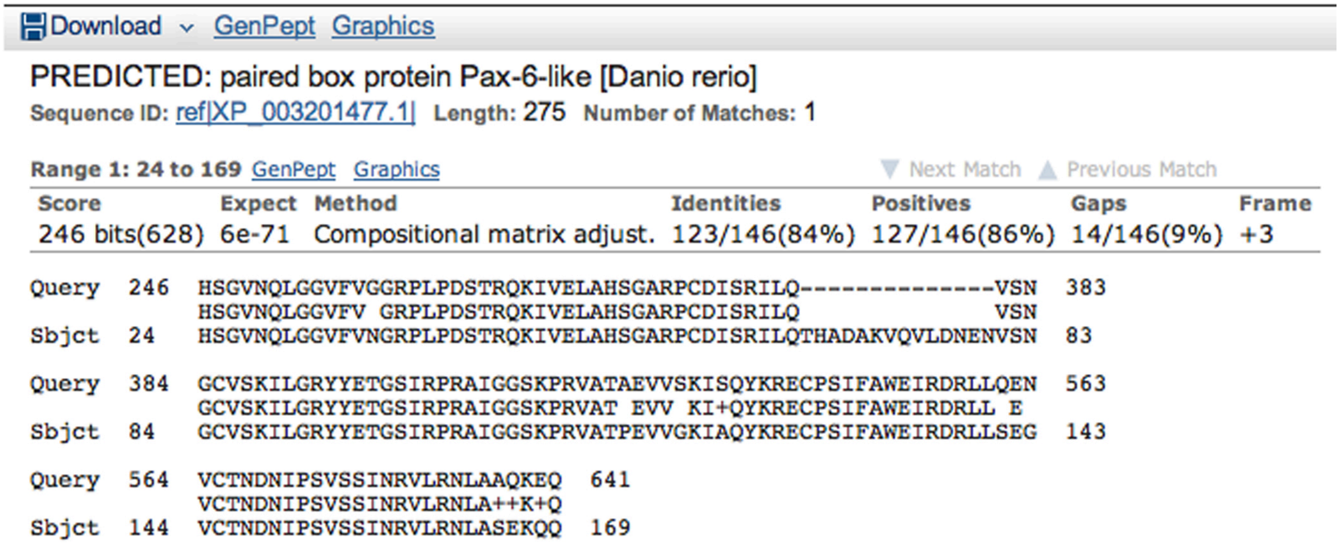
☑ Show results in a new window

We have used a six-frame translation of the query sequence to search a protein sequence databases. Five of the translations are not meaningful. Three of them translate the wrong strand of the sequence. Of the three that translate the correct strand, one of them is the biologically meaningful translation in the correct reading frame, while the other two translate the wrong reading frame.

There is a new component to the results, shown on the next page. The protein sequence database has identified conserved domains in this protein, sequences that show up in related proteins in *Drosophila melanogaster* (paralogs), and in related proteins in other species (orthologs and paralogs). The top of the visual display shows the location of these conserved domains in the protein sequence.


Show Conserved Domains
Putative conserved domains have been detected, click on the image below for detailed results.
1
500
1000
1500
2000
2500
2904
RF +3
DNA binding site
specific DNA base contacts
DNA binding site
Specific hits
PAX
homeodoma
Superfamilies
PAX superfamily
homeodoma
Multi-domains
PAX
Distribution of 167 Blast Hits on the Query Sequence
Mouse over to see the defline, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
550
1100
1650
2200
2750

In the descriptions, click the link to the *Danio rerio* (zebrafish) alignment, shown below.



PREDICTED: paired box protein Pax-6-like [Danio rerio]

Sequence ID: ref|XP_003201477.1| Length: 275 Number of Matches: 1

Range 1: 24 to 169 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps	Frame
246 bits(628)	6e-71	Compositional matrix adjust.	123/146(84%)	127/146(86%)	14/146(9%)	+3

```
Query  246  HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSN  383
            HSGVNQLGGVFV GRPLPDSTRQKIVELAHSGARPCDISRILQ              VSN
Sbjct  24   HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQTHADAKVQVLDNENVSN  83

Query  384  GCVSKILGRYYETGSIRPRAIGGSKPRVATAEVVSKISQYKRECPSIFAWEIRDRLLQEN  563
            GCVSKILGRYYETGSIRPRAIGGSKPRVAT EVV KI+QYKRECPSIFAWEIRDRLL E
Sbjct  84   GCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVGKIAQYKRECPSIFAWEIRDRLLSEG  143

Query  564  VCTNDNIPSVSSINRVLRNLAAQKEQ  641
            VCTNDNIPSVSSINRVLRNLA++K+Q
Sbjct  144  VCTNDNIPSVSSINRVLRNLASEKQQ  169
```

Notice that the alignment is good but not perfect. There is an E value of 6e-71, which is highly significant. BLAST has introduced a gap into the query sequence to improve the alignment. There are parts of the alignment where the amino acid in the *D. melanogaster* protein does not match the amino acid in the *D. rerio* protein.

Notice when there is a mismatch, the program either leaves a blank space in the middle line (nonconservative substitution), or places a "+" sign in the middle line (conservative substitution). The nonconservative substitutions are:

G → N	Glycine to Asparagine
A → P	Alanine to Proline
S → G	Serine to Glycine
Q → S	Glutamine to Serine
N → G	Asparagine to Glycine

The conservative substitutions are:

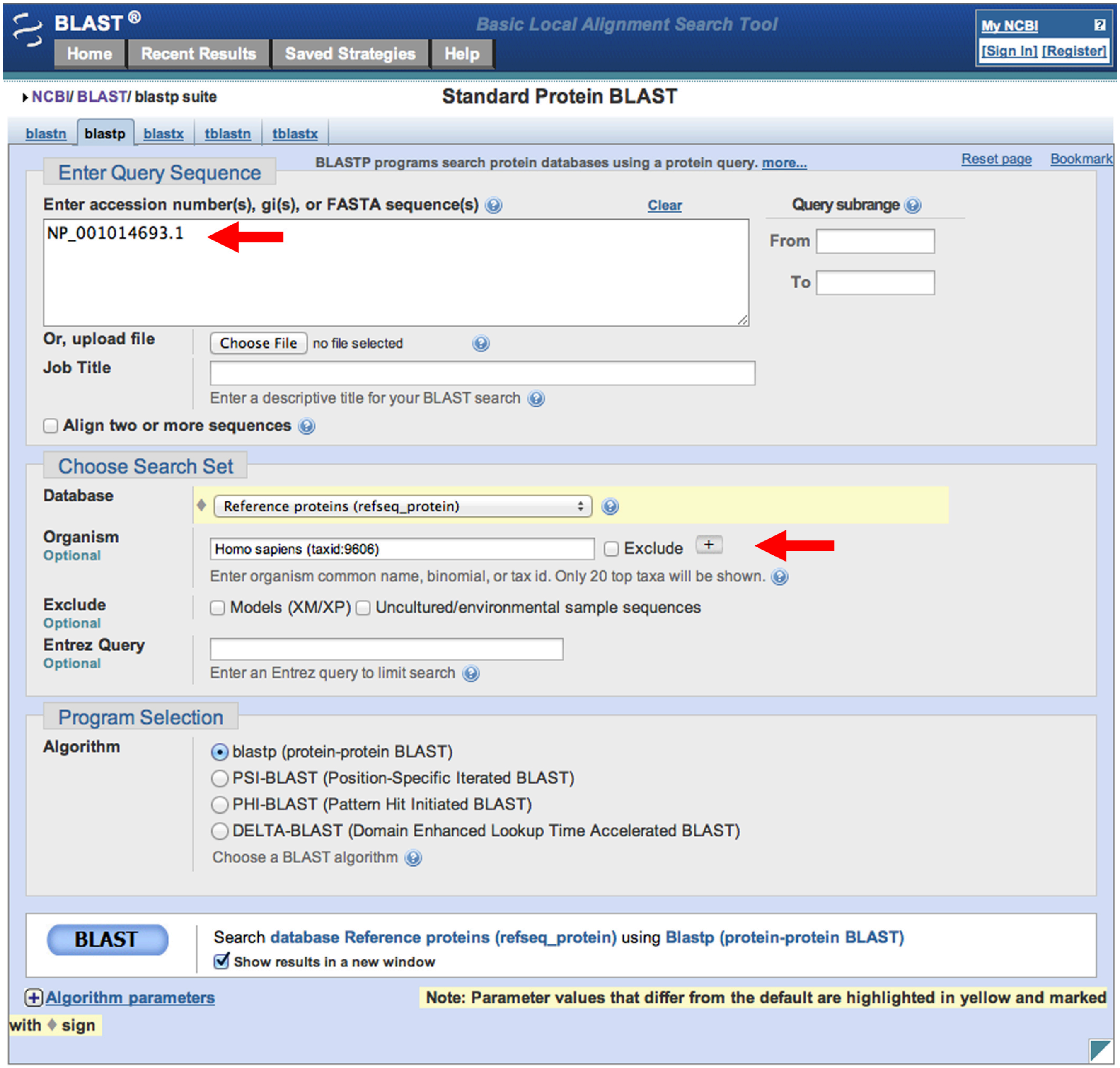
S → A	Serine to Alanine
A → S	Alanine to Serine
Q → E	Glutamine to Glutamic Acid
E → Q	Glutamic Acid to Glutamine

We can discuss what makes some amino acid substitutions conservative and others nonconservative in class.

4. Search for a homologous protein in humans

The top hit in the last BLAST search is the Drosophila protein with the accession ID NP_001014693.1. Copy the accession ID or the complete protein sequence and go to the blastp page.

Enter **NP_001014693.1** or the protein sequence in the search box, restrict the species to **Homo sapiens** and click **BLAST**.



Standard Protein BLAST

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

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s): NP_001014693.1

Query subrange: From ___ To ___

Or, upload file: Choose File no file selected

Job Title: Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Reference proteins (refseq_protein)

Organism (Optional): Homo sapiens (taxid:9606) □ 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

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)

Choose a BLAST algorithm

BLAST — Search database Reference proteins (refseq_protein) 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

What kind of results did you get? What do you think that it means?