BLASTing Through the Kingdom of Life

5/28/2010

Chris Overall
Department of Bioinformatics & Genomics
University of North Carolina - Charlotte

BLASTing Through the Kingdom of Life

Learning Objectives

- Learn about sequence alignment.
  - What is sequence alignment?
  - What is biological sequence alignment?
  - Why is biological sequence alignment important?
- Learn how to use BLAST for biological sequence alignment.

BLASTing Through the Kingdom of Life

What is a Sequence Alignment?

- Arrange two or more words (sequences of letters) so that you can identify regions of similarity between them.
- For the current discusion, the alphabet (set of letters) is the standard English alphabet (size = 26).
- Consider these four words: bat, hat, batter, hatter.
- How similar are bat and hat?
- Whic word is more similar to bat: hat or batter? What about hatter? Why?
- Which pair of words is more similar to each other: (bat, hat) or (batter, hatter)? Why?

BLASTing Through the Kingdom of Life

Sequence Alignments → Longest Common Subsequence

- Let's be a bit more formal about determining similarity.
- For a particular alignment, we will score them as follows:
  - Perfect match = 1 (A – A, B – B, etc.)
  - Mismatch = 0 (A – B, B – A, etc.)
- Using this scoring scheme, we'll find the longest common subsequence in each word pair and score them to determine the most similar sequences.

BLASTing Through the Kingdom of Life

Sequence Alignments → Longest Common Subsequence Examples

Alignment 1

```
BAT
 ||
HAT
```

Score = 2

Alignment 2

```
HAT
 ||
BATTER
```

Score = 2

Alignment 3

```
BAT
 ||
HATTER
```

Score = 2

Alignment 4

```
BAT
|||
BATTER
```

Score = 3

Alignment 5

```
HAT
|||
HATTER
```

Score = 3

Alignment 6

```
BATTER
 |||||
HATTER
```

Score = 5

BLASTing Through the Kingdom of Life

Sequence Alignments → Gapped Alignments

- Consider these three sequences:
  1. THE CATS IN THE HAT
  2. THE CAT IN THE HAT
  3. THE CAT IS A HAT
- Which phrase (2 or 3) is the most similar to phrase 1?

BLASTing Through the Kingdom of Life

Sequence Alignments → Gapped Alignments

- Again, let's be a bit more formal about determining similarity.
- For a particular alignment, we will score it as follows:
  - Perfect match = 1 (A – A, B – B, etc.)
  - Mismatch = 0 (A – B, B – A, etc.)
  - Gap open = -1
  - Gap extension = 0
- Using this scoring scheme, we'll find the longest gapped subsequence in each word pair and score them to determine the most similar sequences.

BLASTing Through the Kingdom of Life

Sequence Alignments → Gapped Alignment Examples

Alignment 1

```
THECATSINTHEHAT
|||||| 
THECATINTHEHAT
```

Score = 6

Alignment 2

```
THECATSINTHEHAT
|||||| |||||||||
THECAT-INTHEHAT
```

Score = 14 – 1 = 13

Alignment 3

```
THECATSINTHEHAT
|||||| 
THECATISAHAT
```

Score = 6

Alignment 4

```
THECATSINTHEHAT
|||||||     |||
THECAT---ISAHAT
```

Score = 6 - 1 + 3 = 8

BLASTing Through the Kingdom of Life

Sequence Alignments → Global vs. Local Alignment

- Global alignment → find the single best alignment across the entire length of both sequences.
- Local alignment → find one or more highly similar local regions between both sequences.

BLASTing Through the Kingdom of Life

What is a Biological Sequence Alignment?

- Arrange two or more molecular sequences (DNA, RNA, protein) to identify regions of similarity between them.
  - Matches, mismatches and gaps.
  - Pairwise vs. multiple sequence alignment (MSA)
- We'll focus on DNA.
- DNA is a sequence of four possible nucleotides: adenine, guanine, cytosine and thymine.
  - DNA alphabet (letters) = {A, C, G, T}
- Much smaller than English alphabet.

BLASTing Through the Kingdom of Life

Biological Sequence Alignment Example

- How would you globally align these two sequences?
  - Sequence 1 → ACTG
  - Sequence 2 → ACGTG

BLASTing Through the Kingdom of Life

Biological Sequence Alignment Example

Alignment 1

```
ACTG-
 | |
ACGTG
```

Alignment 2

```
-ACTG
   | |
ACGTG
```

Alignment 3

```
AC-TG
|| ||
ACGTG
```

BLASTing Through the Kingdom of Life

Why are Biological Sequence Alignments Important?

- The more similar two molecular sequences are, the more likely that the molecules are also similar in structure, function or evolutionary history.
- DNA sequences
  - Mismatch = point mutation
  - Gap = insertion or deletion (indel)
- When developing primers for a PCR reaction (you'll learn about this later), need to make sure that the primer is unique to the organism's genome.

BLASTing Through the Kingdom of Life

Why Do We Need Computers and Algorithms to Find Alignments?

- Why not find all biological sequence alignments manually?
- How many times can you find the query sequence ATCGGCCATTAC in the following target sequence? Is it there at all? If so, is it unique?

```
ATCACTGTAGTAGTAGCTGGAAAGAGAAATCTGTGACTCCAATTAGCCAGTTCCTGCAGACCTTGTGAGGACTAG
AGGAAGAATGCTCCTGGCTGTTTTGTACTGCCTGCTGTGGAGTTTCCAGACCTCCGCTGGCCATTTCCCTAGAGC
CTGTGTCTCCTCTAAGAACCTGATGGAGAAGGAATGCTGTCCACCGTGGAGCGGGGACAGGAGTCCCTGTGGCCA
GCTTTCAGGCAGAGGTTCCTGTCAGAATATCCTTCTGTCCAATGCACCACTTGGGCCTCAATTTCCCTTCACAGG
GGTGGATGACCGGGAGTCGTGGCCTTCCGTCTTTTATAATAGGACCTGCCAGTGCTCTGGCAACTTCATGGGATT
CAACTGTGGAAACTGCAAGTTTGGCTTTTGGGGACCAAACTGCACAGAGAGACGACTCTTGGTGAGAAGAAACAT
CTTCGATTTGAGTGCCCCAGAGAAGGACAAATTTTTTGCCTACCTCACTTTAGCAAAGCATACCATCAGCTCAGA
CTATGTCATCCCCATAGGGACCATTGGCCAAATGAAAAATGGATCAACACCCATGTTTAACGACATCAATATTTA
TGACCTCTTTGTCTGGATGCATTATTATGTGTCAATGGATGCACTGCTTGGGGGATCTGAAATCTGGAGAGACAT
TGATTTTGCCCATGAAGCACCAGCTTTTCTGCCTTGGCATAGACTCTTCTTGTTGCGGTGGGAACAAGAAATCCA
GAAGCTGACAGGAGATGAAAACTTCACTATTCCATATTGGGACTGGCGGGATGCAGAAAAGTGTGACATTTGCAC
AGATGAGTACATGGG
```

BLASTing Through the Kingdom of Life

What is BLAST?

- Basic Local Alignment Search Tool
- Very fast tool for finding local regions of similarity between sequences.
  - Given a query sequence, search one or more target databases for matches.
- Target databases are extremely large; millions of sequences.
  - nt = non-redundant nucleotide sequence database
  - nr = non-redundant protein sequence database

BLASTing Through the Kingdom of Life

Interpreting BLAST Results

- For a given query sequence, you will receive a list of matching target sequences with corresponding alignments.
  - The query sequence can match more than one region for a target sequence (local alignments).
- Each alignment will have an E-value; statistical significance of alignment.
  - Number of sequence matches you would expect to find in a target database composed entirely of random sequences.

Interpreting BLAST Results

- Low E-value means that the match is not random.
- What does an E-value = 4 mean?
- A couple observations about E-values to remember:
  - Greater chance of finding a random match in a larger target database (larger E-value).
  - Greater chance of finding a random match for a shorter query sequence (larger E-value).

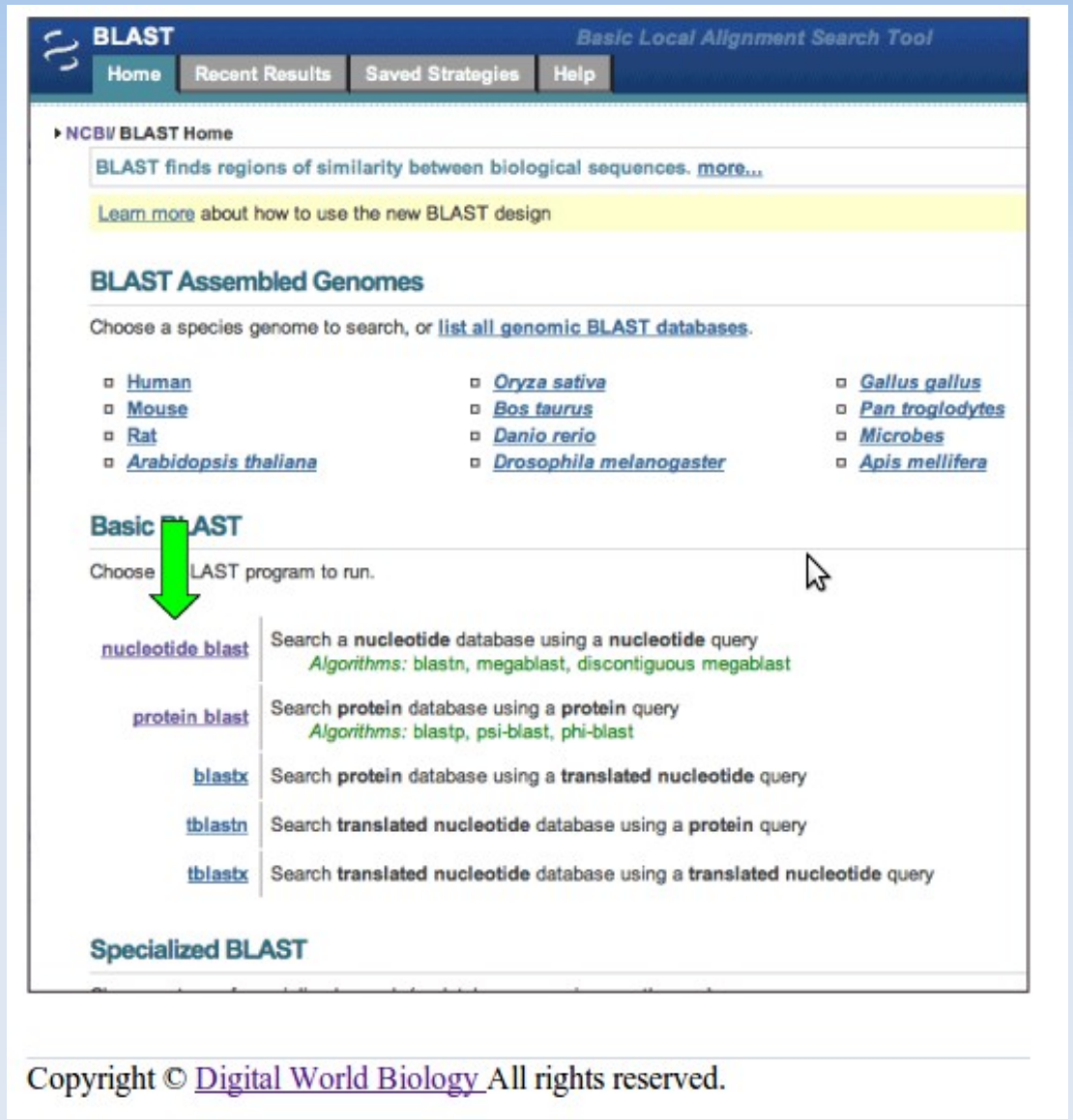
BLASTing Through the Kingdom of Life

Hands-On Tutorial

- Go to http://blast.ncbi.nlm.nih.gov/Blast.cgi

BLASTing Through the Kingdom of Life

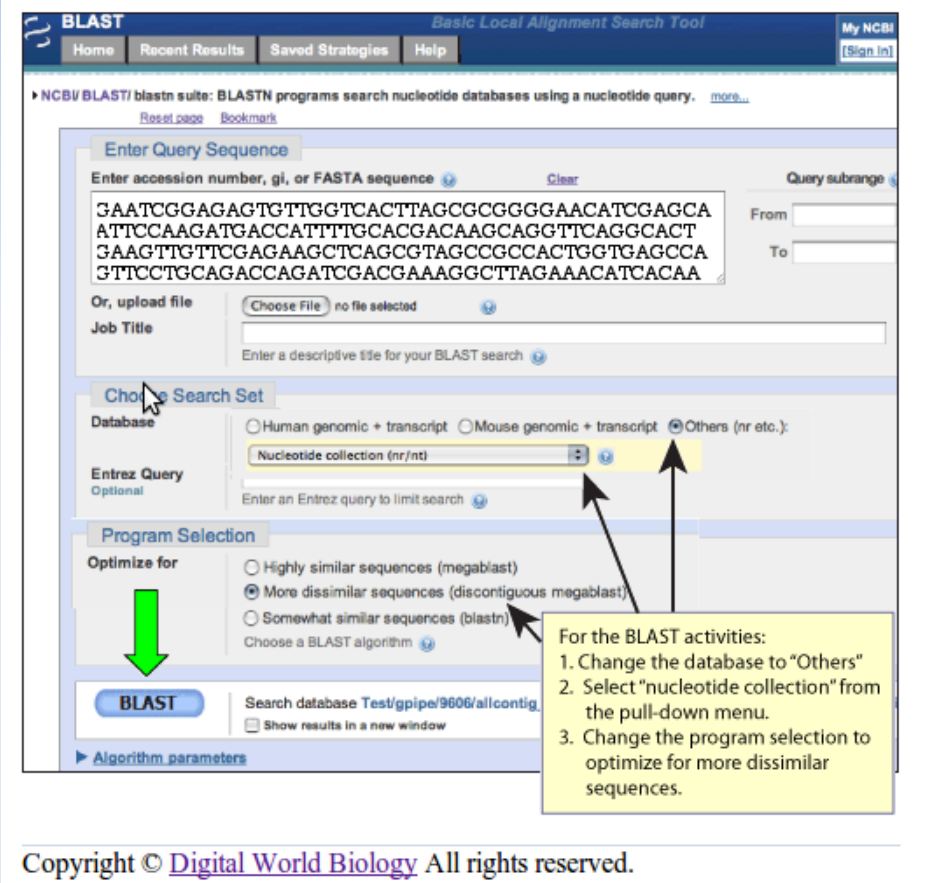
Hands-On Tutorial



BLASTing Through the Kingdom of Life

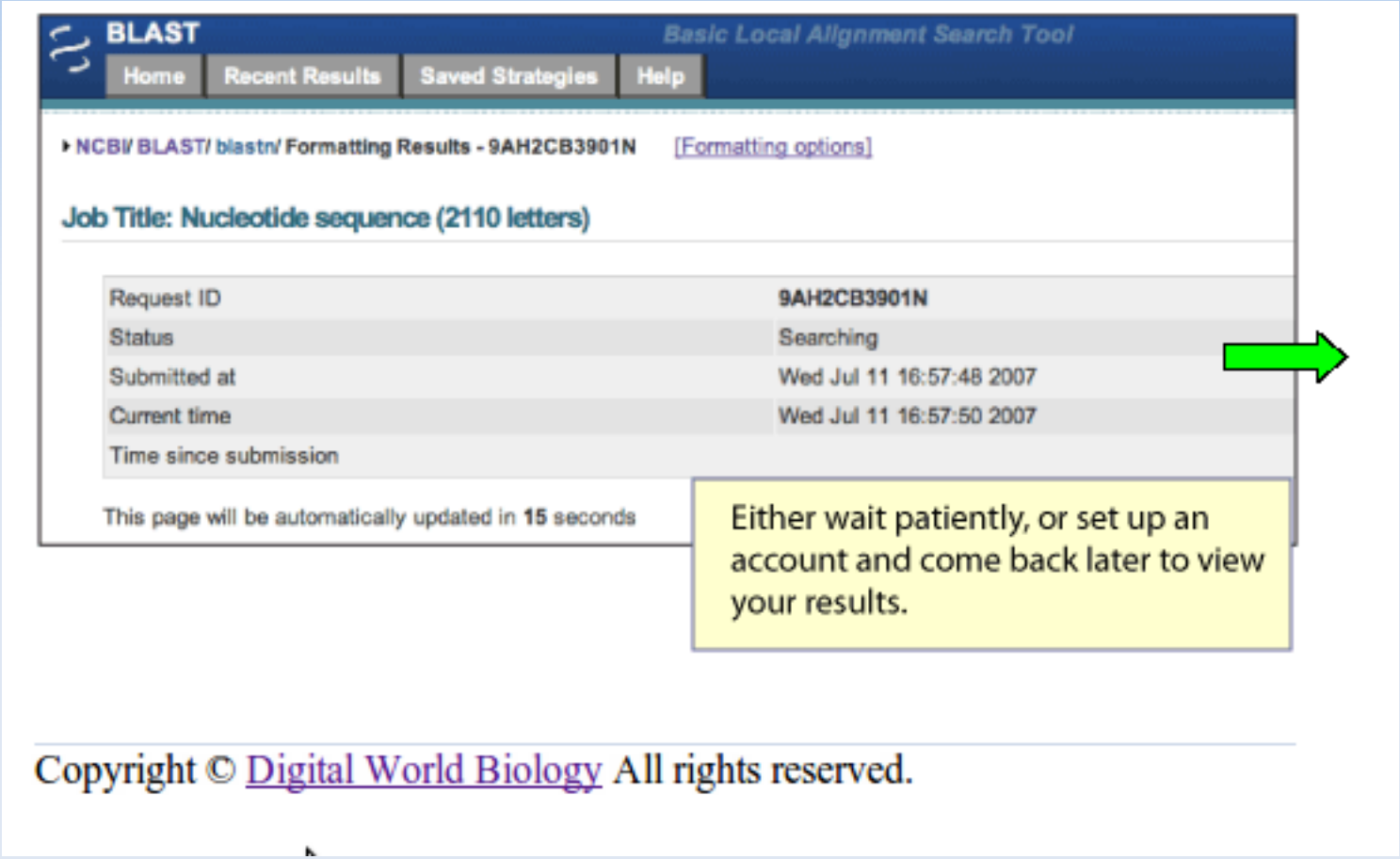
Hands-On Tutorial

- Copy the first query sequence (> Example) from ***http://www.digitalworldbiology.com/BLAST/62000sequences.html***
- Paste target sequence into textbox on BLAST start screen.
- Additional step:
  - Algorithm parameters → Set max target sequences to 500.



BLASTing Through the Kingdom of Life

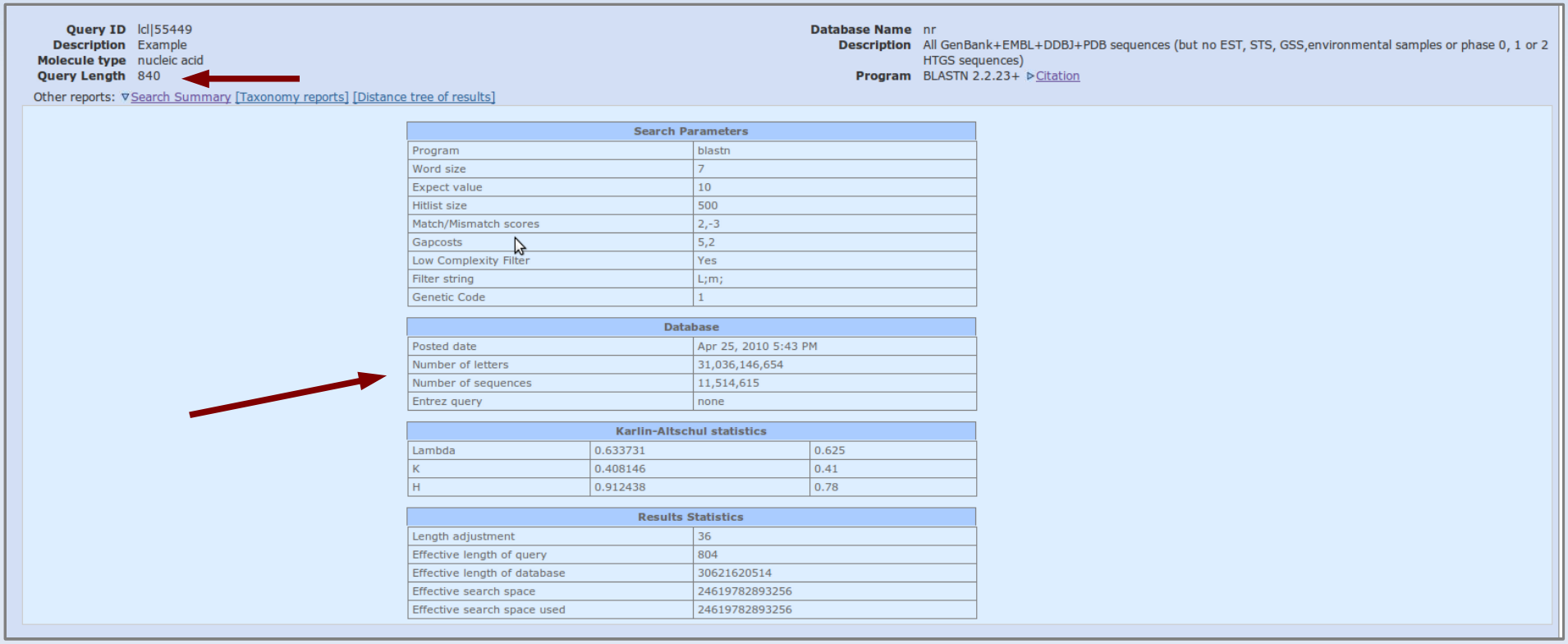
Hands-On Tutorial



BLASTing Through the Kingdom of Life

Hands-On Tutorial

- How many sequences are in the database?
- How many letters are in the database?
- How long is the query sequence?



Query ID lcl|55449
Description Example
Molecule type nucleic acid
Query Length 840

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 2.2.23+ ▹Citation

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

Search Parameters	
Program	blastn
Word size	7
Expect value	10
Hitlist size	500
Match/Mismatch scores	2,-3
Gapcosts	5,2
Low Complexity Filter	Yes
Filter string	L;m;
Genetic Code	1

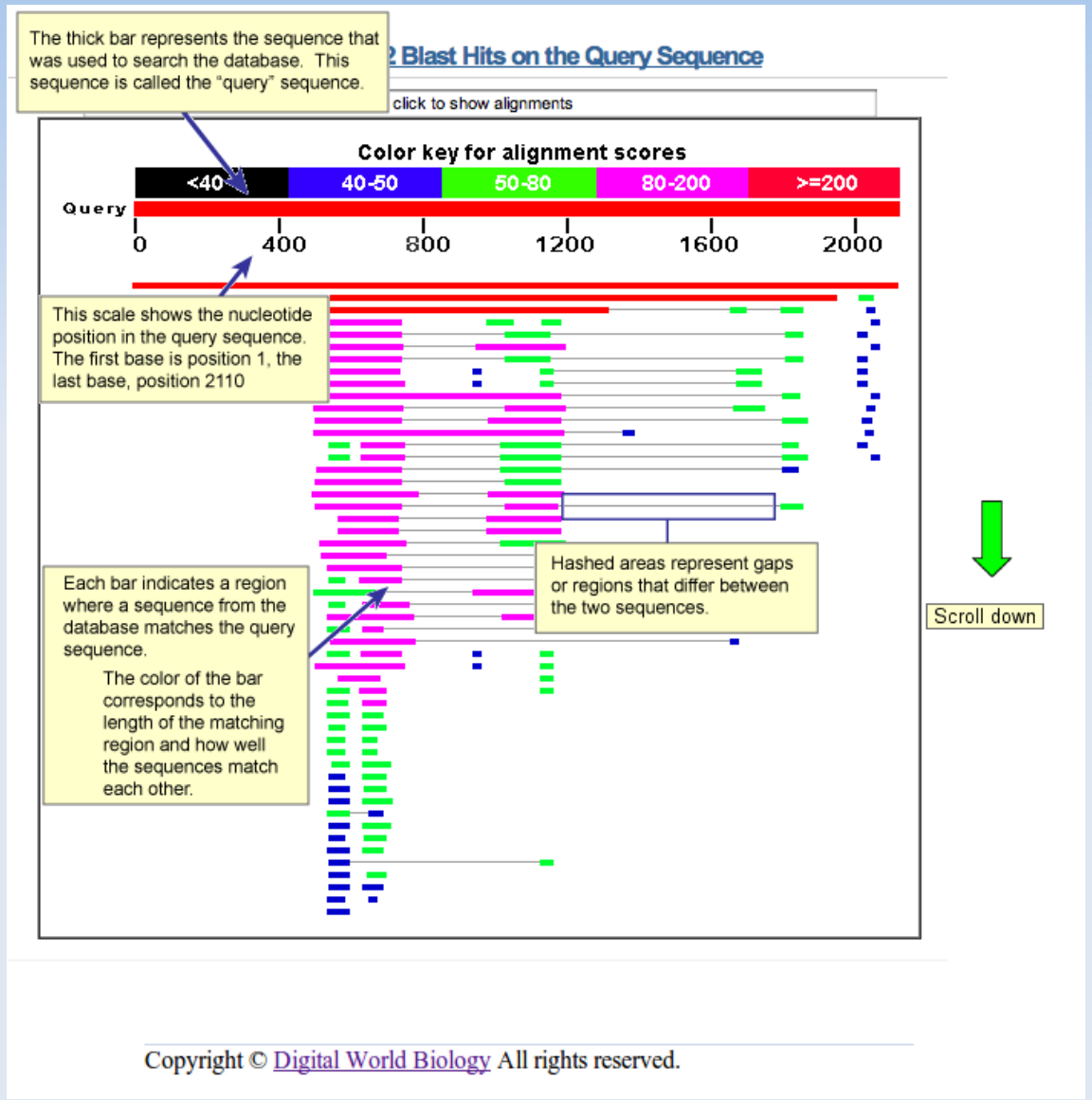
Database	
Posted date	Apr 25, 2010 5:43 PM
Number of letters	31,036,146,654
Number of sequences	11,514,615
Entrez query	none

Karlin-Altschul statistics		
Lambda	0.633731	0.625
K	0.408146	0.41
H	0.912438	0.78

Results Statistics	
Length adjustment	36
Effective length of query	804
Effective length of database	30621620514
Effective search space	24619782893256
Effective search space used	24619782893256

BLASTing Through the Kingdom of Life

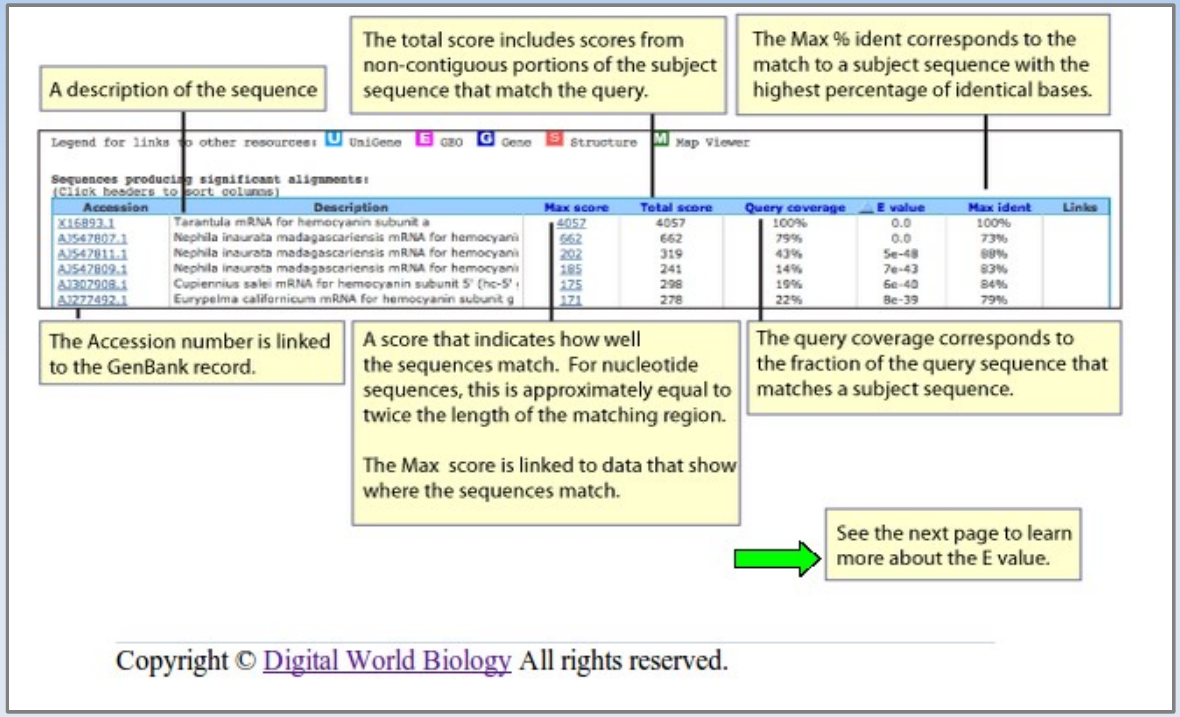
Hands-On Tutorial



BLASTing Through the Kingdom of Life

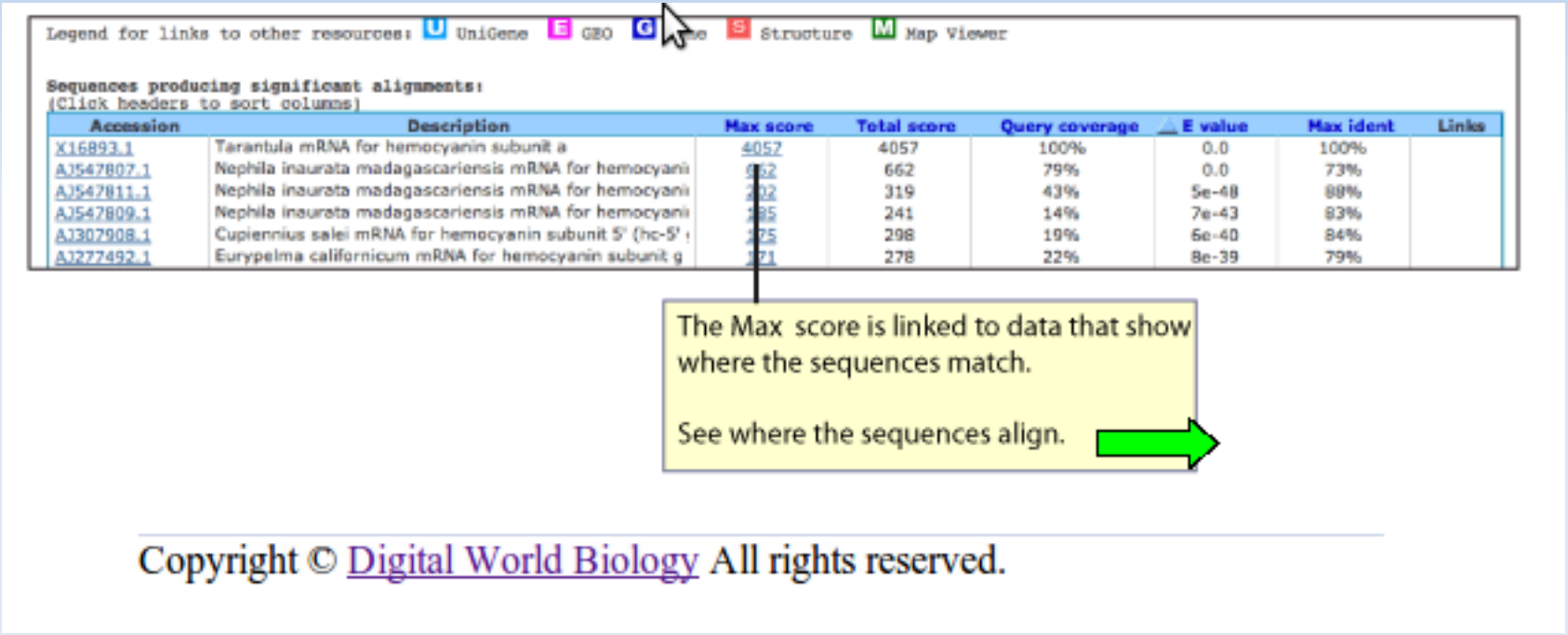
Hands-On Tutorial

- What is the most likely identity of your sequence? What is its accession number?
- What data supports this conclusion?
- What organism is the source of the target sequence?
- Approximately how many alignments have an E-value less than 0.01?
- For which other organisms are there matches?



BLASTing Through the Kingdom of Life

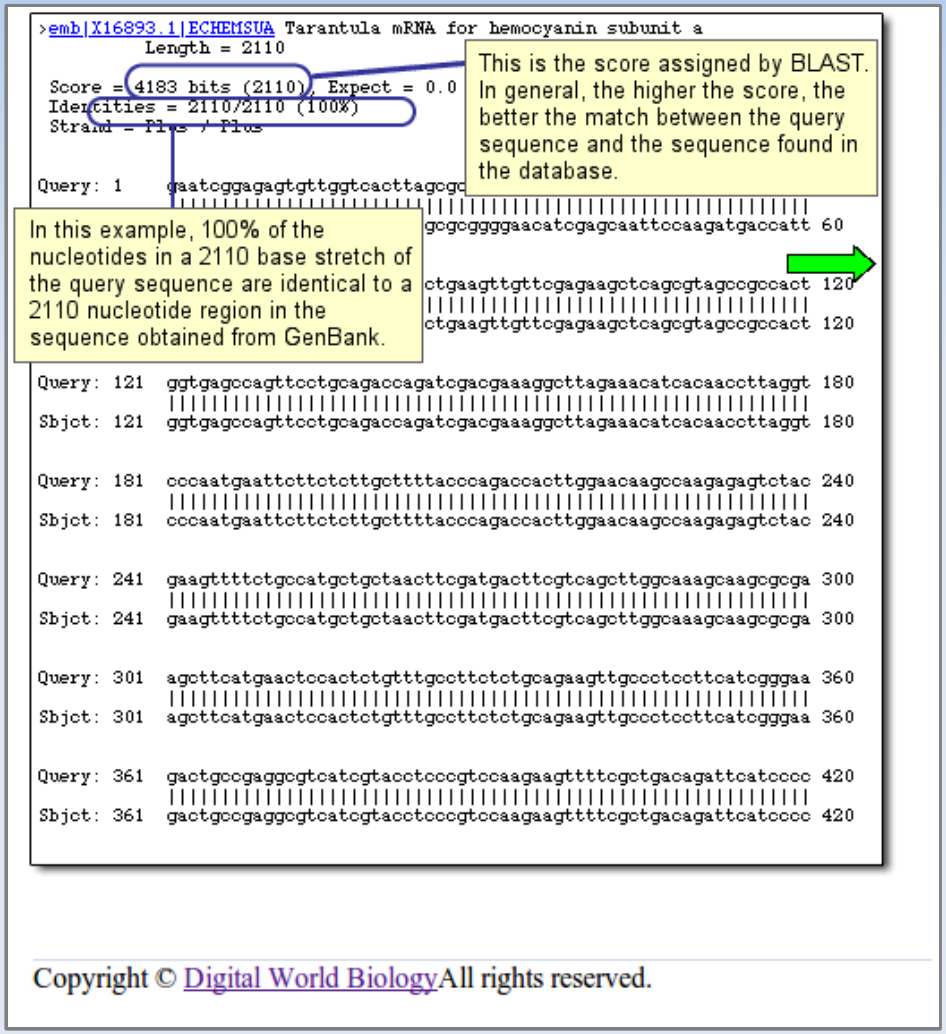
Hands-On Tutorial



BLASTing Through the Kingdom of Life

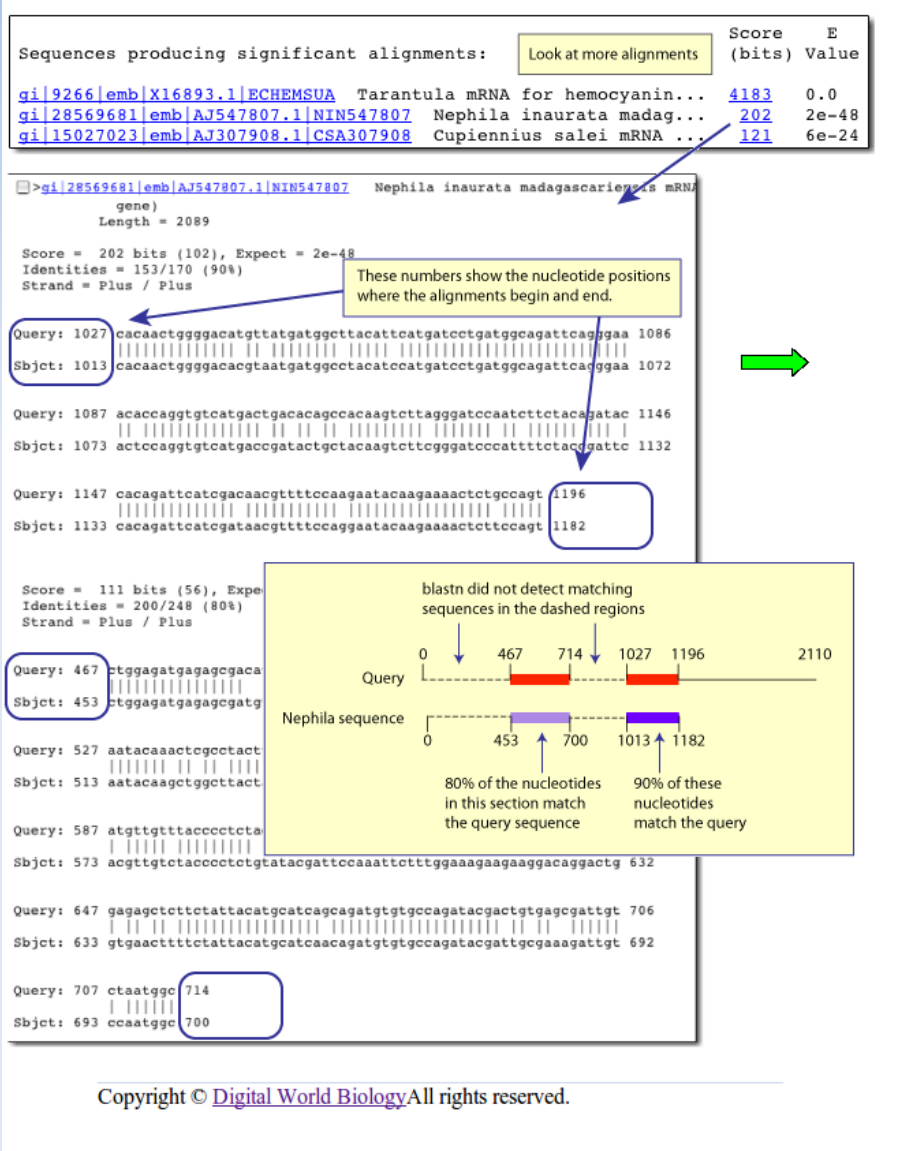
Hands-On Tutorial

- How many identical nucleotides are there for your sequence?



BLASTing Through the Kingdom of Life

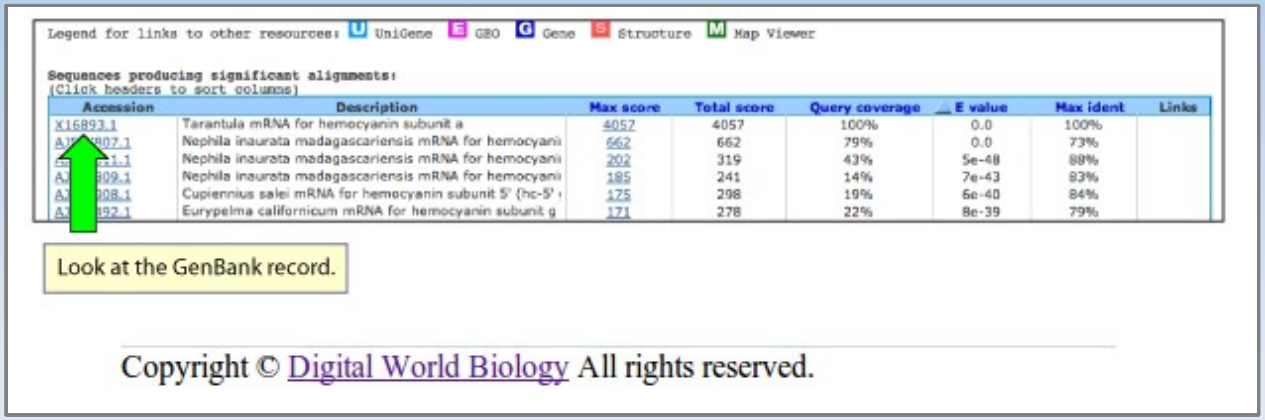
Hands-On Tutorial



BLASTing Through the Kingdom of Life

Hands-On Tutorial

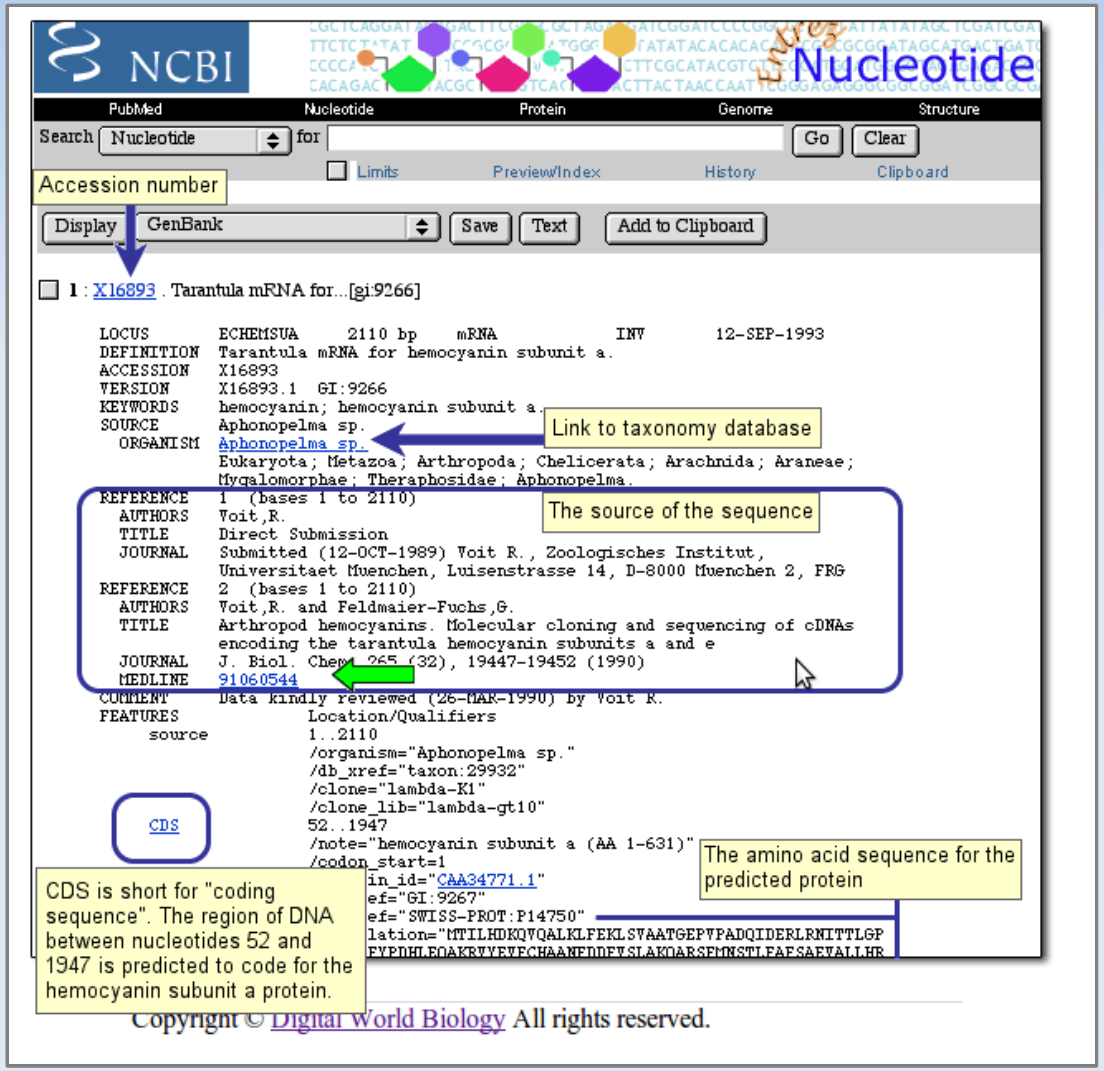
- Find the sequence ”NM_000372.4”
- Is your sequence expressed?
- How do you know?
- How is this sequence related to the first alignment in the results?



BLASTing Through the Kingdom of Life

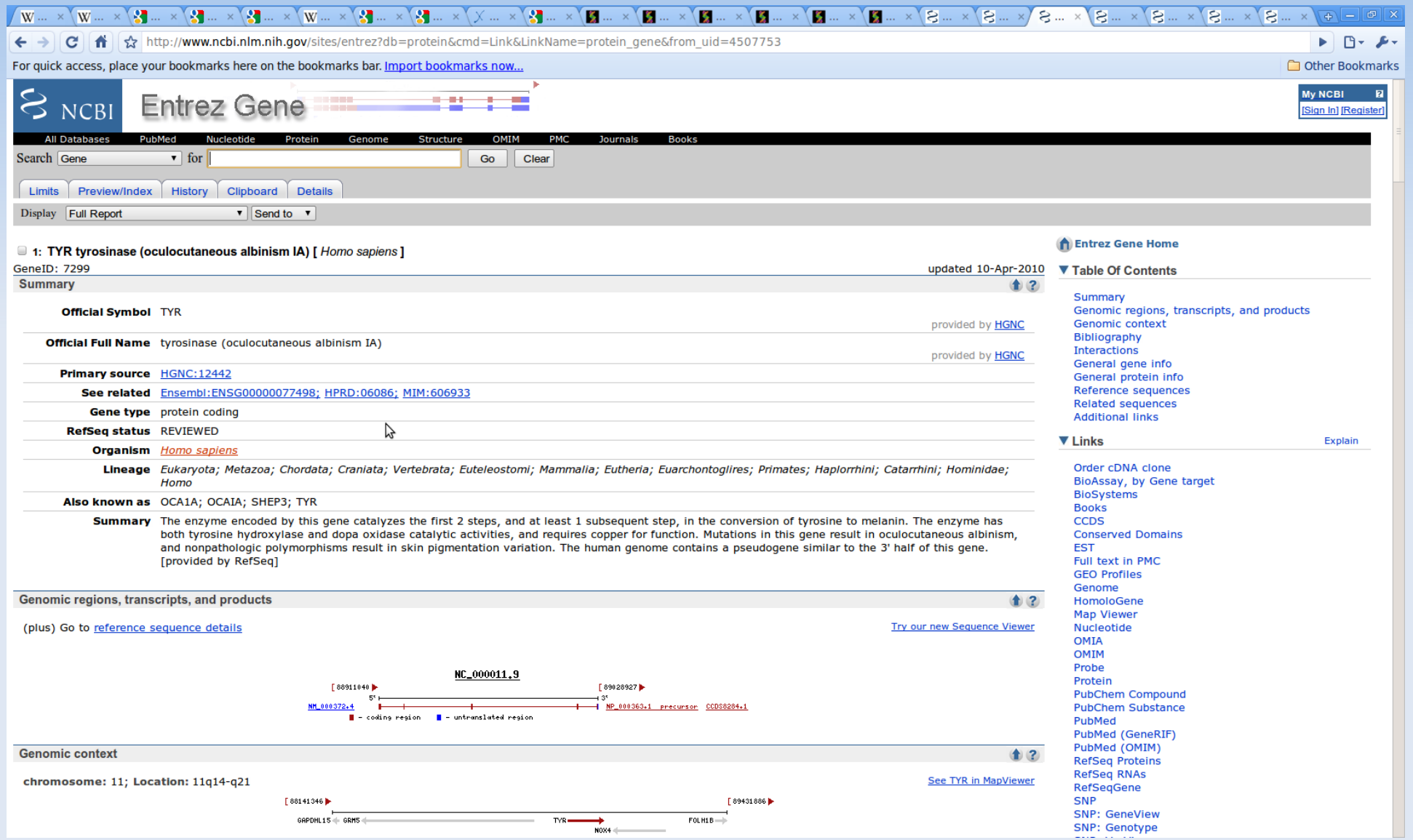
Hands-On Tutorial

- On which chromosome is the matching target sequence located?



BLASTing Through the Kingdom of Life

Hands-On Tutorial



BLASTing Through the Kingdom of Life

Hands-On Tutorial