Introduction to Sequence Alignment

Chris Overall

Department of Bioinformatics and Genomics

University of North Carolina – Charlotte

Sequence alignment algorithms

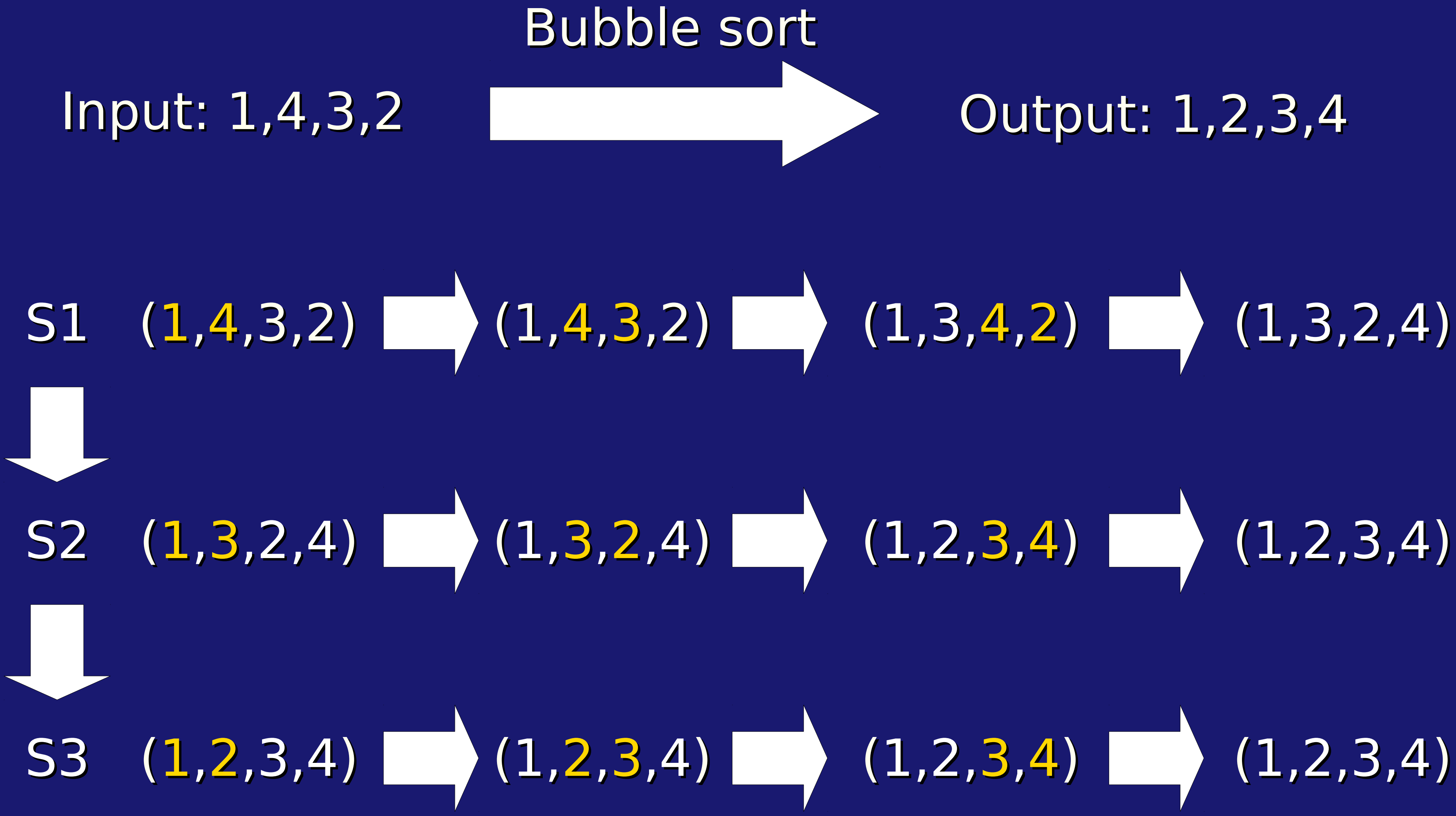
An algorithm is a step-by-step procedure for solving a problem or accomplishing something

Examples

- Finding prime numbers
- Sorting numbers from lowest to highest
- Calculating shortest delivery routes
- A cooking recipe
- A lab protocol

The term algorithm usually refers to solving a problem (often mathematical) using a computer

Bubble sort is a classic algorithm for sorting numbers



Bubble sort

Input: 1,4,3,2

Output: 1,2,3,4

S1 (1,4,3,2) (1,4,3,2) (1,3,4,2) (1,3,2,4)

S2 (1,3,2,4) (1,3,2,4) (1,2,3,4) (1,2,3,4)

S3 (1,2,3,4) (1,2,3,4) (1,2,3,4) (1,2,3,4)

<u>Sequence alignment algorithms</u> are ways to arrange two or more “words” to find out how similar they are to each other

Word

Sequence of letters (symbols) using a particular alphabet (set of letters)

Doesn't have to be the English alphabet

Examples

FRAGILE

SUPERCALIFRAGILISTICEXPIALIDOCIOUS

1324647597

? 🎁 ✕ ■ ? ✈ ⓘ ✦ 🚓 ◀ ⏩ 🚓 ✔ □ ⊘

Let's look at a simple example using words from the English alphabet (size = 26)

Consider these four words: hat, bat, hatter, batter

How similar are bat and hat? Why?

Which word is more similar to bat? Why?

- hat
- batter

Which pair of words is more similar to each other? Why?

- (bat, hat)
- (batter, hatter)

One way to figure out how similar two words (sequences) are is to find the longest common subsequence

What's the algorithm to do this?

Step 1: Left-align the two sequences

Step 2: Score the alignment

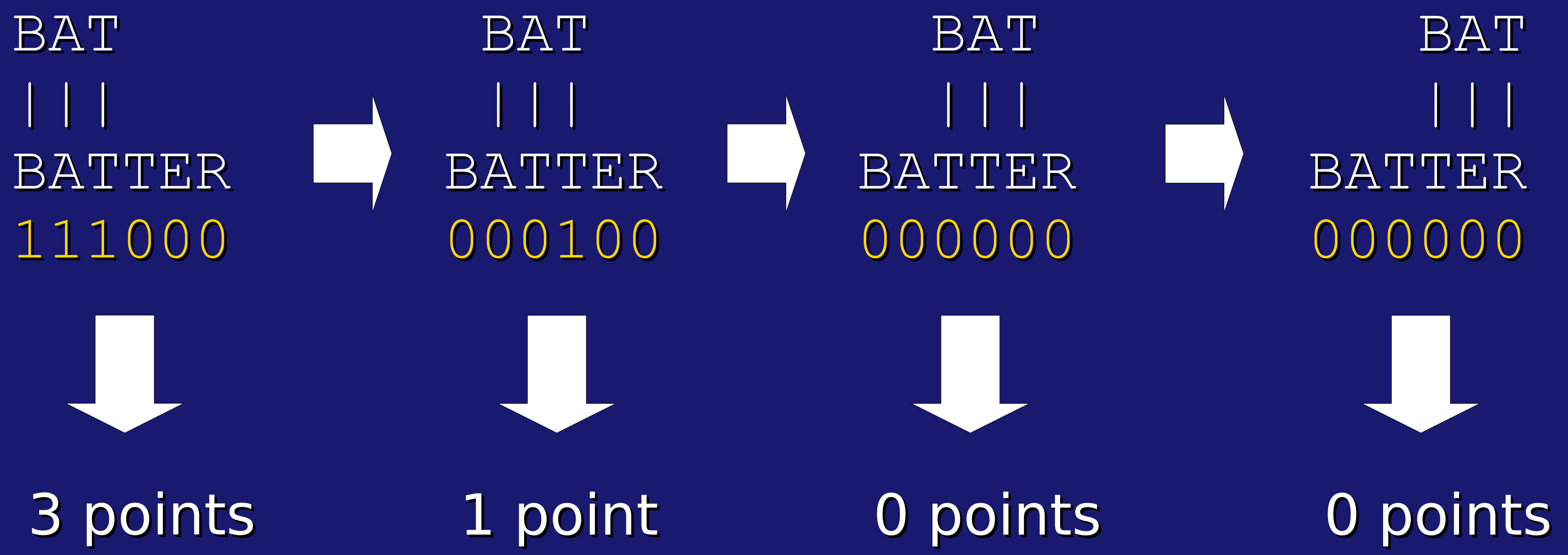
Step 3: Shift the shorter alignment to the right by one character and repeat steps 1-2 until you reach the end of the longer sequence

Here's a simple example of finding the longest common subsequence

We'll evaluate the sub-alignments using this scoring scheme

Match = 1 point

Mismatch = 0 point



```
BAT          BAT          BAT          BAT
|||          |||          |||          |||
BATTER      BATTER      BATTER      BATTER
111000      000100      000000      000000
```

3 points

1 point

0 points

0 points

Now we can score the alignments between our four words: hat, hatter, bat, batter

Alignment 1

```
BAT
|||
HAT
011
```

Score = 2

Alignment 2

```
HAT
|||
BATTER
011
```

Score = 2

Alignment 3

```
BAT
|||
HATTER
011
```

Score = 2

Alignment 4

```
BAT
|||
BATTER
111000
```

Score = 3

Alignment 5

```
HAT
|||
HATTER
111000
```

Score = 3

Alignment 6

```
BATTER
||||||
HATTER
011111
```

Score = 5

Just when you thought you understood alignments, it gets a bit more complicated

Consider these three phrases ➡

THE CATS IN THE HAT

THE CAT IN THE HAT

THE CAT IS A HAT

Which two phrases are the most similar?

What algorithm did you use to figure that out?

Will our previous alignment method work?
Why or why not?

The current algorithm can't finding LCSs with additional or missing characters (insertions and deletions)

We'll use the same scoring scheme as before

Match = 1 point

Mismatch = 0 points

```
THECATINTHEHAT
||||||||||||||
THECATSINTHEHAT
11111100000000
```

6 points

```
 THECATINTHEHAT
 ||||||||||||||
THECATSINTHEHAT
 00000011111111
```

8 points

How can we fix the algorithm?

We can improve the algorithm by allowing gaps in the longest common subsequence

We'll use a modified scoring scheme →

- Match = 1
- Mismatch = 0
- Gap open = -1
- Gap extension = 0

```
THECAT-INTHEHAT
|||||| ||||||||
THECATSINTHEHAT
111111-11111111
```

→ 6 - 1 + 8 = 13

Better!

You might be asking yourselves, why do we need the gap penalty?

Which of these two alignments is better? Why?

```
THECATINTHEHAT
||||||||||||||
THECATINTHEHAT

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

Without a gap penalty, both alignments have the same score (14)

We need gap penalties to reflect the intuition that, all things being equal, ungapped alignments are better than gapped alignments

Now we can score the alignments between our three phrases (sequences)

Alignment 1

```
THECAT-INTHEHAT
|||||| ||||||||
THECATSINTHEHAT
111111-11111111
```

Score = 13

Alignment 2

```
THECAT---ISAHAT
||||||   ||||||
THECATSINTHEHAT
111111-00000111
```

Score = 8

Alignment 3

```
THECATI--SAHAT
|||||||  |||||
THECATINTHEHAT
1111111-000111
```

Score = 9

Alternative alignments for 2

Alignment 2a

```
THECAT---ISAHAT
||||||   ||||||
THECATSINTHEHAT
111111-00000111
```

Score = 8

Alignment 2b

```
THECAT-I--SAHAT
|||||| |  |||||
THECATSINTHEHAT
111111-1-000111
```

Score = 8

Biological sequence alignments

Biological sequence alignment algorithms are ways of arranging two or more molecular sequences to identify regions of similarity between them

Types of molecules →

- DNA
- RNA
- Protein

We'll focus on DNA, which consists of four nucleotides (alphabet size = 4) →

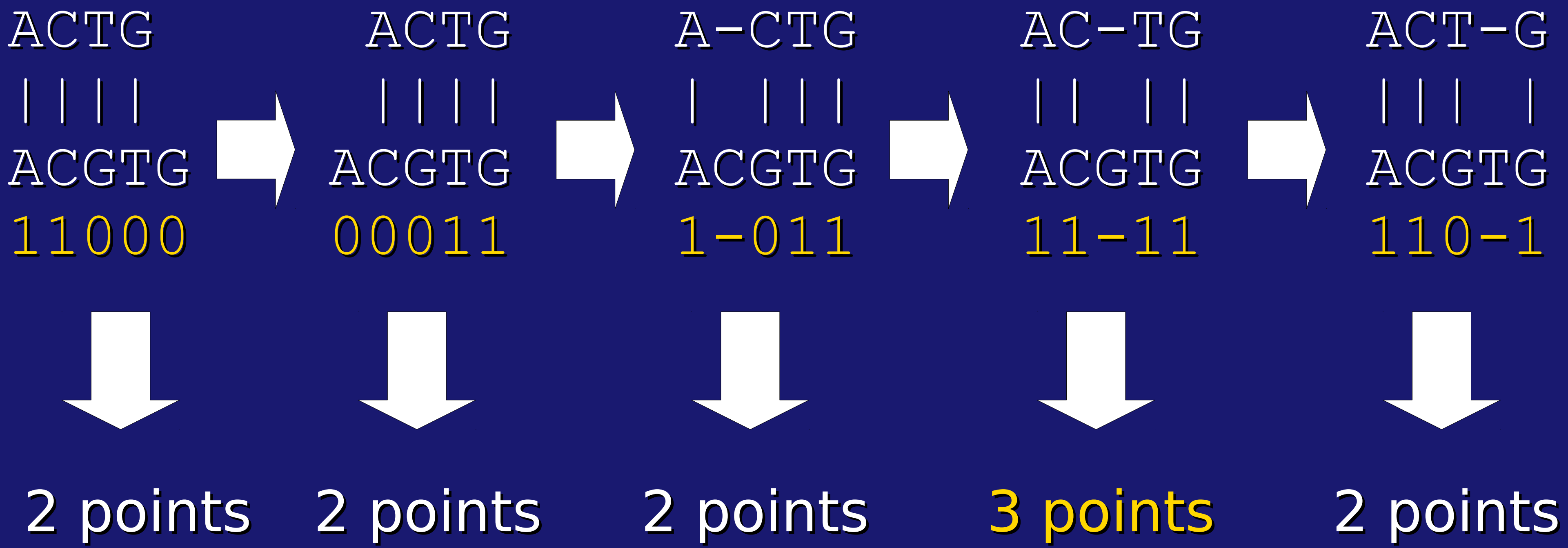
- Adenine (A)
- Thymine (T)
- Guanine (G)
- Cytosine (C)

We align biological sequences in the same way as we did with English words and phrases

Example → aligning two DNA sequences

ACTG vs. ACGTG

Match = 1, mismatch = 0, gap open = -1, gap extension = 0



```
ACTG
||||
ACGTG
11000
```

2 points

```
 ACTG
 ||||
ACGTG
00011
```

2 points

```
A-CTG
| |||
ACGTG
1-011
```

2 points

```
AC-TG
|| ||
ACGTG
11-11
```

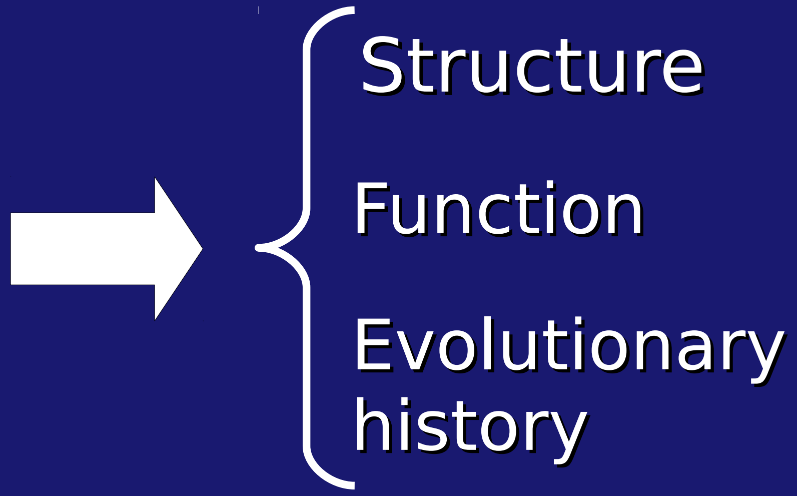
3 points

```
ACT-G
||| |
ACGTG
110-1
```

2 points

Why are biological sequence alignments important?

The more similar two molecular sequences are, the more likely that the molecules are also similar in:

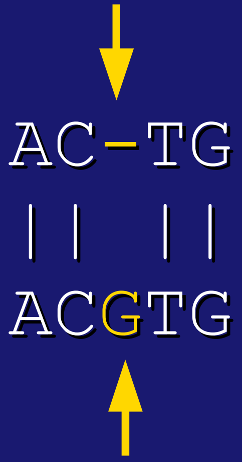


Structure

Function

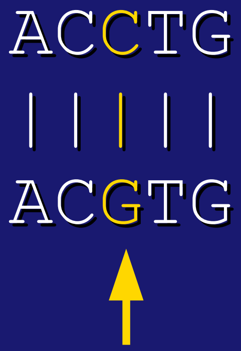
Evolutionary history

Deletion?



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

Insertion?



```
ACCTG
|||||
ACGTG
```

Point mutation

We need computers and algorithms to find biological sequence alignments

Why not find all biological sequence alignments manually?

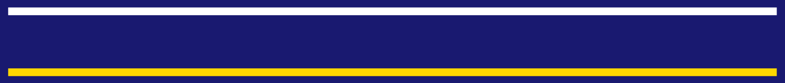
How many times can you find the query sequence ATCGGCCATTAC in the following target sequence? Is it there at all? If so, is it unique?

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

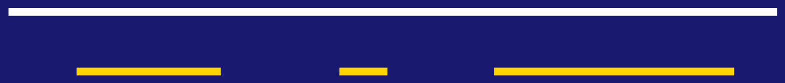
Basic Local Alignment Search Tool (BLAST)

There are two types of sequence alignment: local and global

Global alignment → find the single best alignment across the entire length of both sequences

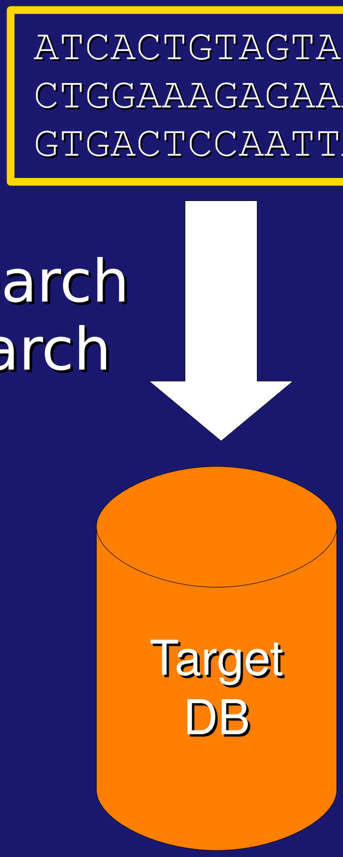


Local alignment → find one or more highly similar local regions between both sequences



BLAST is a very fast tool for finding local regions of similarity between biological sequences

Basic Local Alignment Search Tool



ATCACTGTAGTAGTAG
CTGGAAAGAGAAATCT
GTGACTCCAATTAGCC

Your query sequence

Can be DNA, RNA or protein

Internet search or local search

BLAST your sequence: search target database for local alignments

nt = non-redundant nucleotide sequence database

nr = non-redundant protein sequence database

Target databases are extremely large; millions of sequences

There are six types of BLAST, depending on the type of query and target sequences

Nucleotide BLAST (blastn) → Search a nucleotide database using a nucleotide query

Protein BLAST (blastp) → Search a protein database using a protein query

blastx → Search a protein database using a translated nucleotide query

tblastn → Search a translated nucleotide database using a protein query

tblastx → Search translated nucleotide database using a translated nucleotide query

Many animals use the Earth's magnetic field for orientation and navigation esp. during migration

Some examples: sea turtles, swallows, monarch butterflies and fruit flies (Drosophila melanogaster)

Cryptochome is a key protein for geomagnetic sensing; it seems to be a quantum compass

Humans produce cryptochrome in the retina, but we don't seem to have this geomagnetic perception

Human cryptochrome, when it replaces fruit fly cryptochrome, works the same way

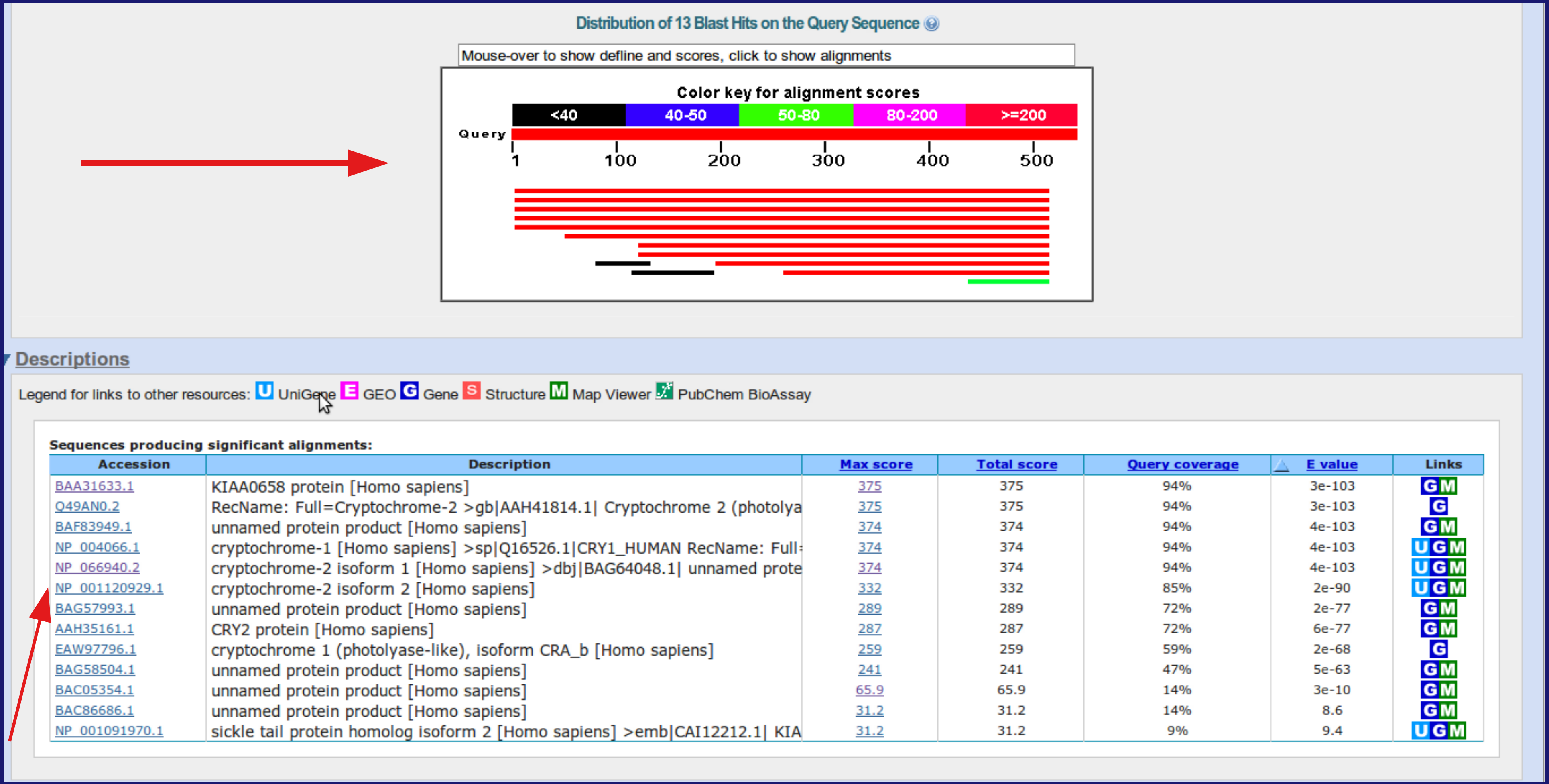
Researchers created cryptochrome-deficient flies, and they lost their ability to navigate

They then created transgenic flies with human cryptochrome instead of their normal version

The flies with the human cryptochrome could navigate just as well as the flies with the normal version

How similar are the protein sequences of human cryptochrome and fly cryptochrome?

We can use BLAST to find protein sequences in humans that are similar to fly cryptochrome



Distribution of 13 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

Descriptions

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

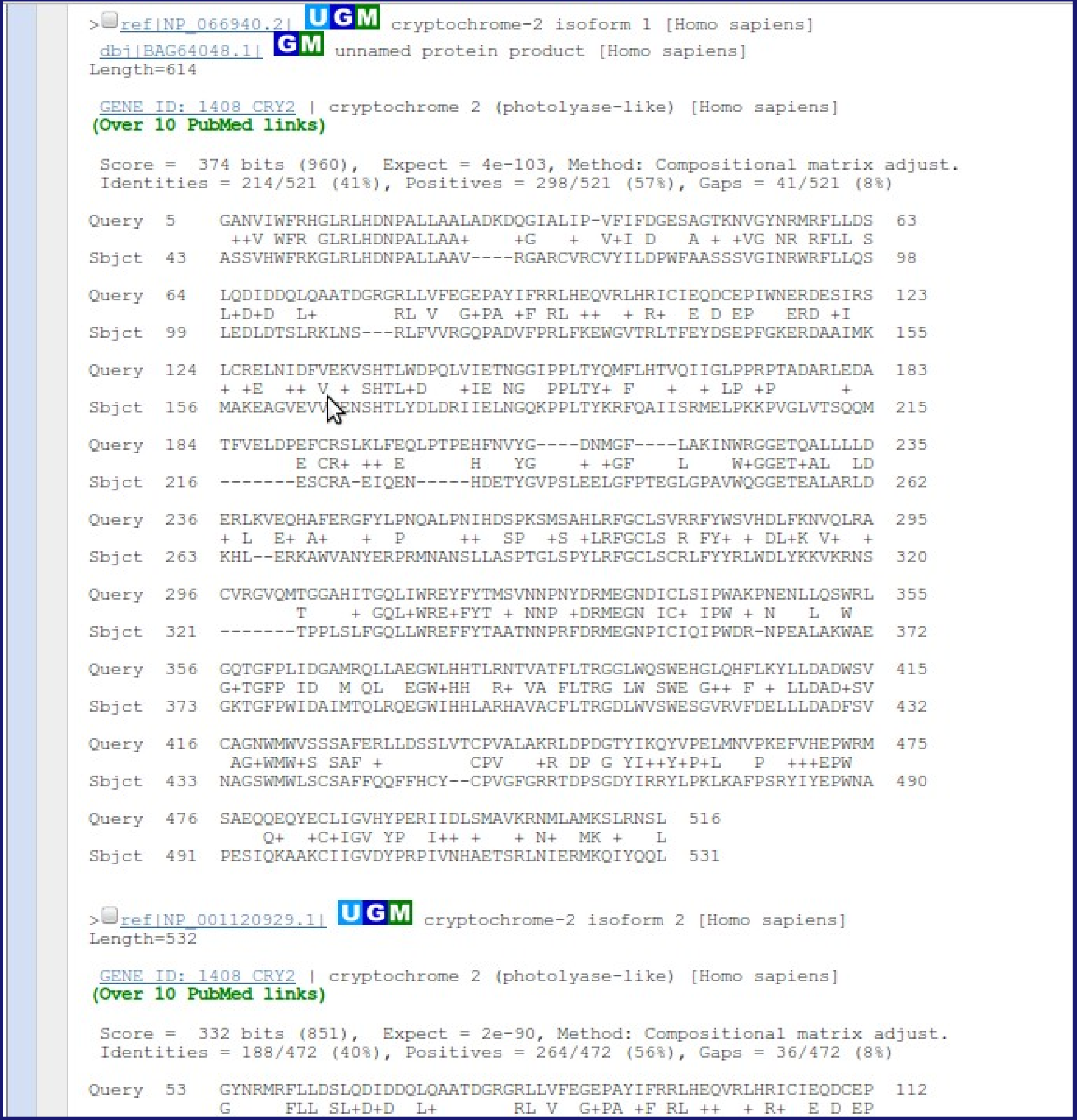
Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
BAA31633.1	KIAA0658 protein [Homo sapiens]	375	375	94%	3e-103	G M
Q49AN0.2	RecName: Full=Cryptochrome-2 >gb|AAH41814.1| Cryptochrome 2 (photolya	375	375	94%	3e-103	G
BAF83949.1	unnamed protein product [Homo sapiens]	374	374	94%	4e-103	G M
NP_004066.1	cryptochrome-1 [Homo sapiens] >sp|Q16526.1|CRY1_HUMAN RecName: Full=	374	374	94%	4e-103	U G M
NP_066940.2	cryptochrome-2 isoform 1 [Homo sapiens] >dbj|BAG64048.1| unnamed prote	374	374	94%	4e-103	U G M
NP_001120929.1	cryptochrome-2 isoform 2 [Homo sapiens]	332	332	85%	2e-90	U G M
BAG57993.1	unnamed protein product [Homo sapiens]	289	289	72%	2e-77	G M
AAH35161.1	CRY2 protein [Homo sapiens]	287	287	72%	6e-77	G M
EAW97796.1	cryptochrome 1 (photolyase-like), isoform CRA_b [Homo sapiens]	259	259	59%	2e-68	G
BAG58504.1	unnamed protein product [Homo sapiens]	241	241	47%	5e-63	G M
BAC05354.1	unnamed protein product [Homo sapiens]	65.9	65.9	14%	3e-10	G M
BAC86686.1	unnamed protein product [Homo sapiens]	31.2	31.2	14%	8.6	G M
NP_001091970.1	sickle tail protein homolog isoform 2 [Homo sapiens] >emb|CAI12212.1| KIA	31.2	31.2	9%	9.4	U G M

We get a list – and visual overview – of alignments of the query sequence to target sequences

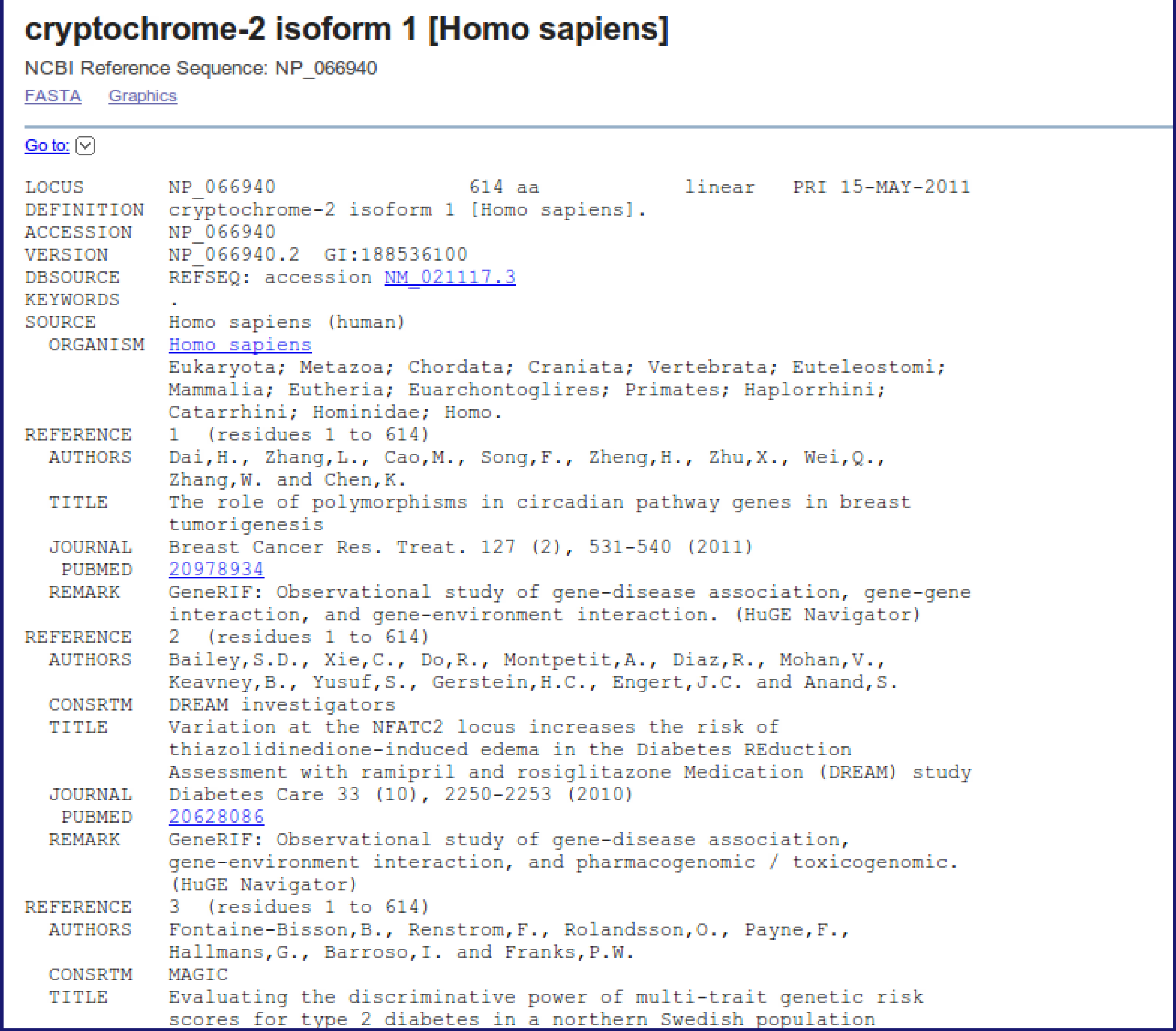
We can view the alignment between query and target sequences for each match

- Query coverage = 94%
- Score = 374 bits (960)
- Expect = 4e-103
- Identities = 214/521 (41%)
- Positives = 298/521 (57%)
- Gaps = 41/521 (8%)



```
>ref|NP_066940.2| UGM cryptochrome-2 isoform 1 [Homo sapiens]
 dbj|BAG64048.1| GM unnamed protein product [Homo sapiens]
Length=614

 GENE ID: 1408 CRY2 | cryptochrome 2 (photolyase-like) [Homo sapiens]
(Over 10 PubMed links)

 Score =  374 bits (960),  Expect = 4e-103, Method: Compositional matrix adjust.
 Identities = 214/521 (41%), Positives = 298/521 (57%), Gaps = 41/521 (8%)

Query  5    GANVIWFRHGLRLHDNPALLAALADKDQGIALIP-VFIFDGESAGTKNVGYNRMRFLLDS  63
             ++V WFR GLRLHDNPALLAA+    +G   +  V+I D   A + +VG NR RFLL S
Sbjct  43   ASSVHWFRKGLRLHDNPALLAAV----RGARCVRCVYILDPWFAASSSVGINRWRFLLQS  98

Query  64   LQDIDDQLQAATDGRGRLLVFEGEPAYIFRRLHEQVRLHRICIEQDCEPIWNERDESIRS  123
            L+D+D  L+       RL V  G+PA +F RL ++  + R+  E D EP   ERD +I
Sbjct  99   LEDLDTSLRKLNS---RLFVVRGQPADVFPRLFKEWGVTRLTFEYDSEPFGKERDAAIMK  155

Query  124  LCRELNIDFVEKVSHTLWDPQLVIETNGGIPPLTYQMFLHTVQIIGLPPRPTADARLEDA  183
            + +E  ++ V + SHTL+D   +IE NG  PPLTY+ F   +  + LP +P       +
Sbjct  156  MAKEAGVEVV[illegible]ENSHTLYDLDRIIELNGQKPPLTYKRFQAIISRMELPKKPVGLVTSQQM  215

Query  184  TFVELDPEFCRSLKLFEQLPTPEHFNVYG----DNMGF----LAKINWRGGETQALLLLD  235
                   E CR+ ++ E        H   YG     + +GF     L    W+GGET+AL  LD
Sbjct  216  -------ESCRA-EIQEN-----HDETYGVPSLEELGFPTEGLGPAVWQGGETEALARLD  262

Query  236  ERLKVEQHAFERGFYLPNQALPNIHDSPKSMSAHLRFGCLSVRRFYWSVHDLFKNVQLRA  295
            + L  E+ A+    +  P      ++  SP  +S +LRFGCLS R FY+ + DL+K V+  +
Sbjct  263  KHL--ERKAWVANYERPRMNANSLLASPTGLSPYLRFGCLSCRLFYYRLWDLYKKVKRNS  320

Query  296  CVRGVQMTGGAHITGQLIWREYFYTMSVNNPNYDRMEGNDICLSIPWAKPNENLLQSWRL  355
                   T    + GQL+WRE+FYT + NNP +DRMEGN IC+ IPW + N   L  W
Sbjct  321  -------TPPLSLFGQLLWREFFYTAATNNPRFDRMEGNPICIQIPWDR-NPEALAKWAE  372

Query  356  GQTGFPLIDGAMRQLLAEGWLHHTLRNTVATFLTRGGLWQSWEHGLQHFLKYLLDADWSV  415
            G+TGFP ID  M QL  EGW+HH  R+ VA FLTRG LW SWE G++ F + LLDAD+SV
Sbjct  373  GKTGFPWIDAIMTQLRQEGWIHHLARHAVACFLTRGDLWVSWESGVRVFDELLLDADFSV  432

Query  416  CAGNWMWVSSSAFERLLDSSLVTCPVALAKRLDPDGTYIKQYVPELMNVPKEFVHEPWRM  475
             AG+WMW+S SAF +          CPV   +R DP G YI++Y+P+L   P  +++EPW
Sbjct  433  NAGSWMWLSCSAFFQQFFHCY--CPVGFGRRTDPSGDYIRRYLPKLKAFPSRYIYEPWNA  490

Query  476  SAEQQEQYECLIGVHYPERIIDLSMAVKRNMLAMKSLRNSL  516
                Q+  +C+IGV YP  I++ +  + N+  MK +   L
Sbjct  491  PESIQKAAKCIIGVDYPRPIVNHAETSRLNIERMKQIYQQL  531


>ref|NP_001120929.1| UGM cryptochrome-2 isoform 2 [Homo sapiens]
Length=532

 GENE ID: 1408 CRY2 | cryptochrome 2 (photolyase-like) [Homo sapiens]
(Over 10 PubMed links)

 Score =  332 bits (851),  Expect = 2e-90, Method: Compositional matrix adjust.
 Identities = 188/472 (40%), Positives = 264/472 (56%), Gaps = 36/472 (8%)

Query  53   GYNRMRFLLDSLQDIDDQLQAATDGRGRLLVFEGEPAYIFRRLHEQVRLHRICIEQDCEP  112
            G     FLL SL+D+D  L+       RL V  G+PA +F RL ++  + R+  E D EP
```

We can also view a detailed record for each matching target sequence



cryptochrome-2 isoform 1 [Homo sapiens]

NCBI Reference Sequence: NP_066940

FASTA Graphics

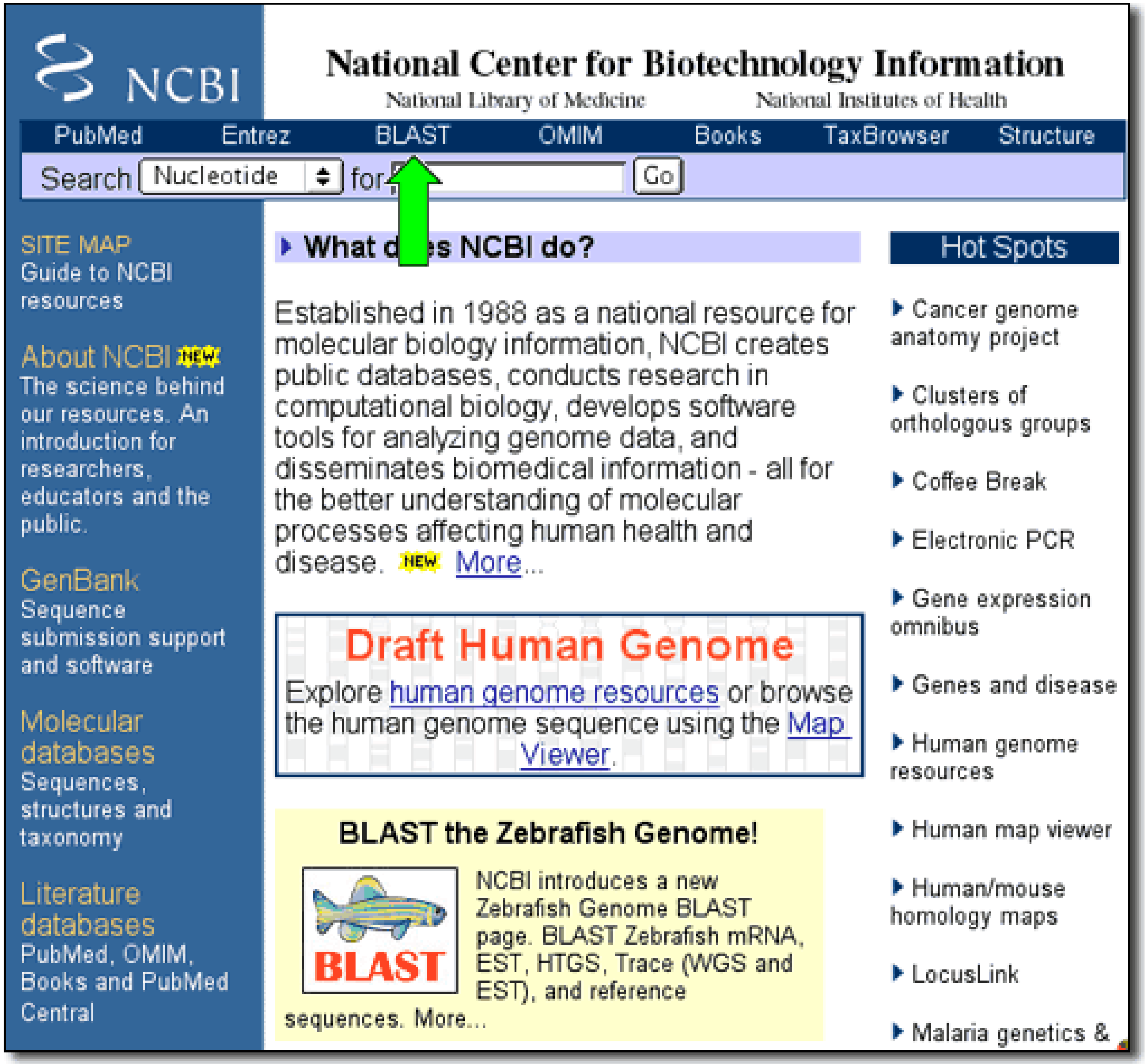
Go to:

```
LOCUS       NP_066940                614 aa            linear   PRI 15-MAY-2011
DEFINITION  cryptochrome-2 isoform 1 [Homo sapiens].
ACCESSION   NP_066940
VERSION     NP_066940.2  GI:188536100
DBSOURCE    REFSEQ: accession NM_021117.3
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 614)
  AUTHORS   Dai,H., Zhang,L., Cao,M., Song,F., Zheng,H., Zhu,X., Wei,Q.,
            Zhang,W. and Chen,K.
  TITLE     The role of polymorphisms in circadian pathway genes in breast
            tumorigenesis
  JOURNAL   Breast Cancer Res. Treat. 127 (2), 531-540 (2011)
   PUBMED   20978934
  REMARK    GeneRIF: Observational study of gene-disease association, gene-gene
            interaction, and gene-environment interaction. (HuGE Navigator)
REFERENCE   2  (residues 1 to 614)
  AUTHORS   Bailey,S.D., Xie,C., Do,R., Montpetit,A., Diaz,R., Mohan,V.,
            Keavney,B., Yusuf,S., Gerstein,H.C., Engert,J.C. and Anand,S.
  CONSRTM   DREAM investigators
  TITLE     Variation at the NFATC2 locus increases the risk of
            thiazolidinedione-induced edema in the Diabetes REduction
            Assessment with ramipril and rosiglitazone Medication (DREAM) study
  JOURNAL   Diabetes Care 33 (10), 2250-2253 (2010)
   PUBMED   20628086
  REMARK    GeneRIF: Observational study of gene-disease association,
            gene-environment interaction, and pharmacogenomic / toxicogenomic.
            (HuGE Navigator)
REFERENCE   3  (residues 1 to 614)
  AUTHORS   Fontaine-Bisson,B., Renstrom,F., Rolandsson,O., Payne,F.,
            Hallmans,G., Barroso,I. and Franks,P.W.
  CONSRTM   MAGIC
  TITLE     Evaluating the discriminative power of multi-trait genetic risk
            scores for type 2 diabetes in a northern Swedish population
```

BLAST Tutorial

http://www.digitalworldbiology.com/BLAST/slide1.html

BLAST Tutorial → Slide 1


NCBI
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
PubMed
Entrez
BLAST
OMIM
Books
TaxBrowser
Structure
Search
Nucleotide
for
Go
Hot Spots
NEW

SITE MAP
Guide to NCBI resources

About NCBI NEW
The science behind our resources. An introduction for researchers, educators and the public.

GenBank
Sequence submission support and software

Molecular databases
Sequences, structures and taxonomy

Literature databases
PubMed, OMIM, Books and PubMed Central

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. NEW More...

Draft Human Genome

Explore human genome resources or browse the human genome sequence using the Map Viewer.

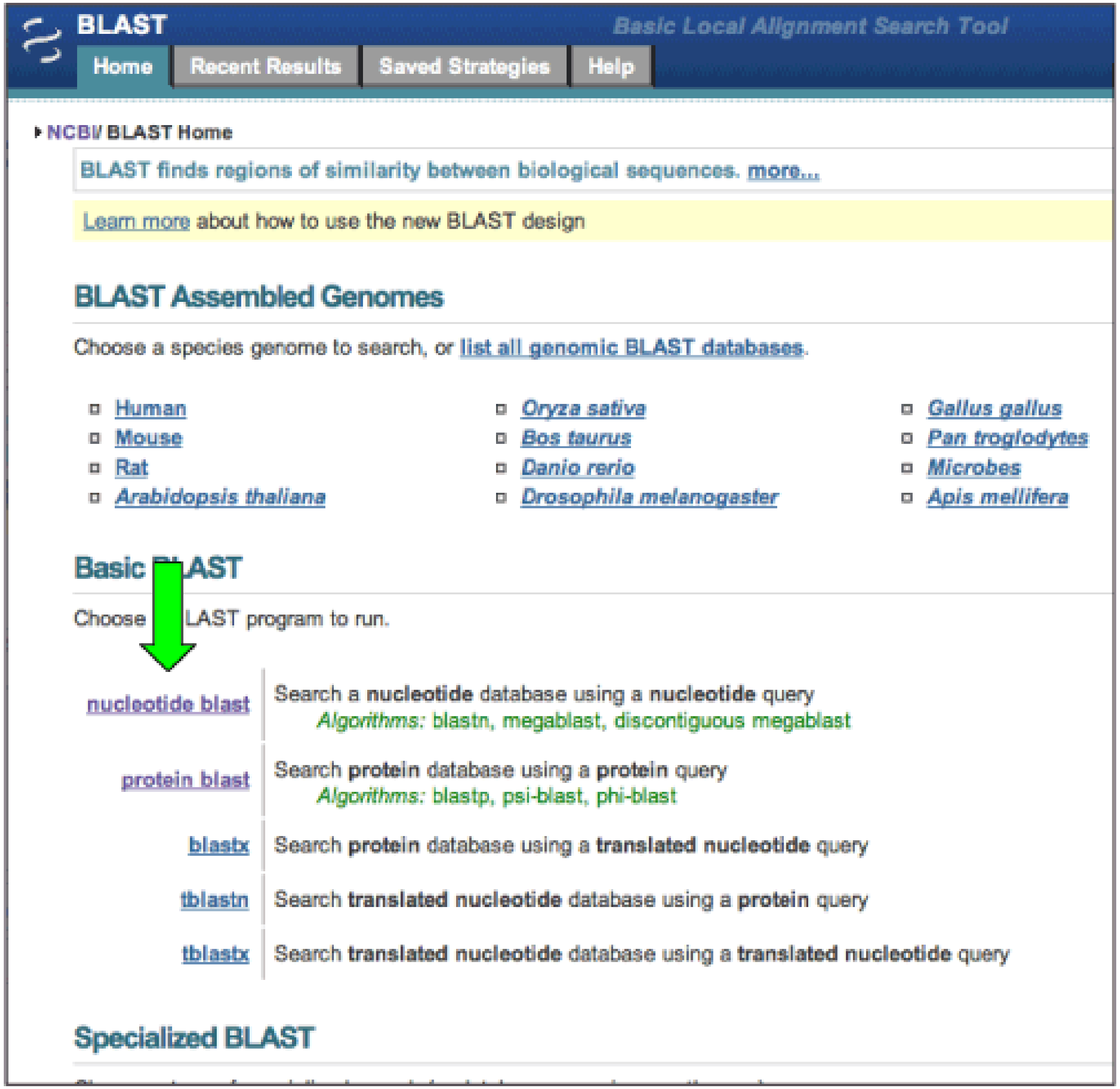
BLAST the Zebrafish Genome!

NCBI introduces a new Zebrafish Genome BLAST page. BLAST Zebrafish mRNA, EST, HTGS, Trace (WGS and EST), and reference sequences. More...

Hot Spots

- Cancer genome anatomy project
- Clusters of orthologous groups
- Coffee Break
- Electronic PCR
- Gene expression omnibus
- Genes and disease
- Human genome resources
- Human map viewer
- Human/mouse homology maps
- LocusLink
- Malaria genetics &

BLAST Tutorial → Slide 2



NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

Learn more about how to use the new BLAST design

BLAST Assembled 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*

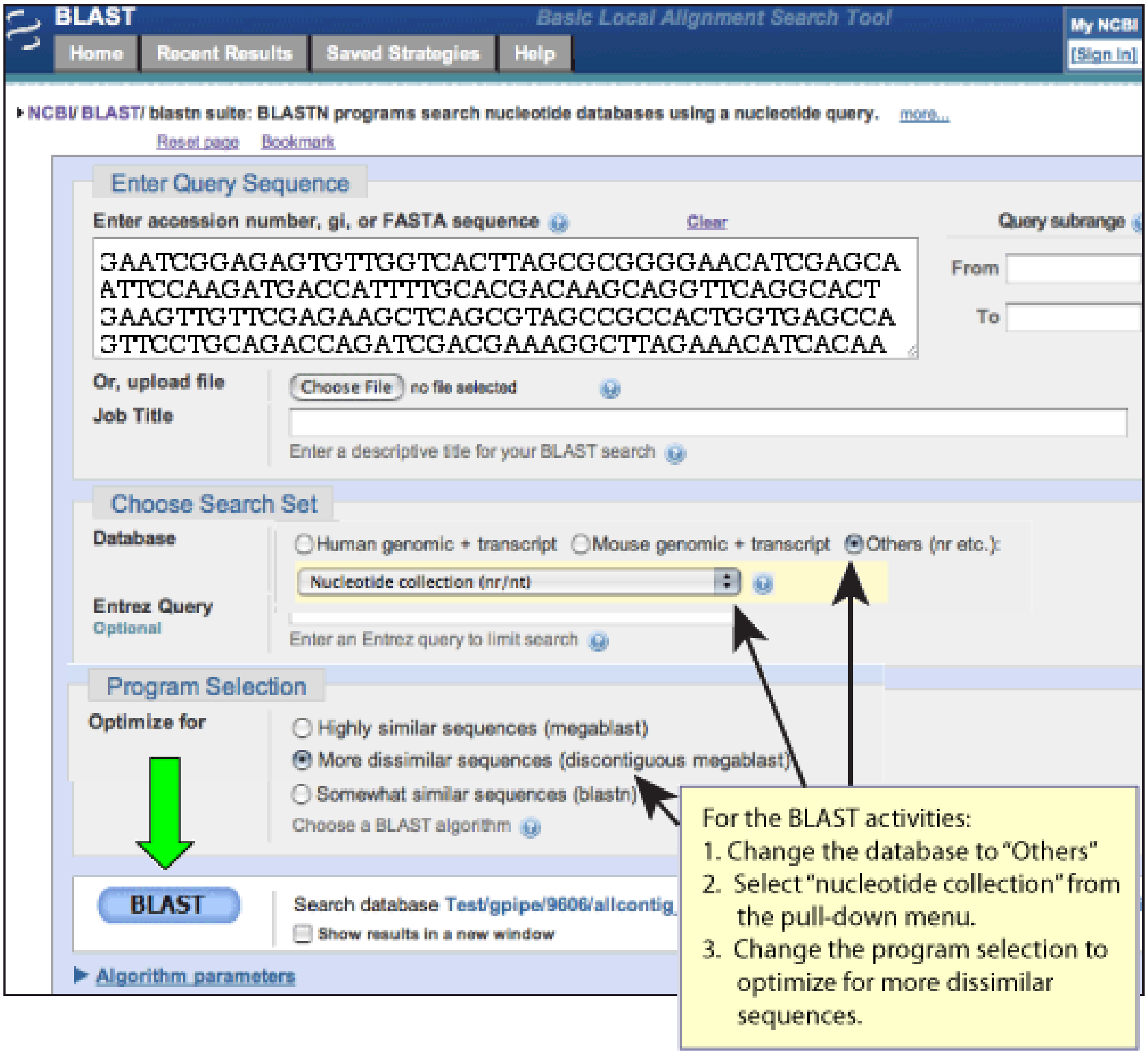
Basic BLAST

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
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

BLAST Tutorial → Slide 3



▸ NCBI/ BLAST/ blastn suite: BLASTN programs search nucleotide databases using a nucleotide query. more...

Reset page Bookmark

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear

```
GAATCGGAGAGTGTTGGTCACTTAGCGCGGGGAACATCGAGCA
ATTCCAAGATGACCATTTTGCACGACAAGCAGGTTCAGGCACT
GAAGTTGTTCGAGAAGCTCAGCGTAGCCGCCACTGGTGAGCCA
GTTCCTGCAGACCAGATCGACGAAAGGCTTAGAAACATCACAA
```

Query subrange

From

To

Or, upload file: Choose File no file selected

Job Title:

Enter a descriptive title for your BLAST search

Choose Search Set

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

Nucleotide collection (nr/nt)

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 Test/gpipe/9606/allcontig

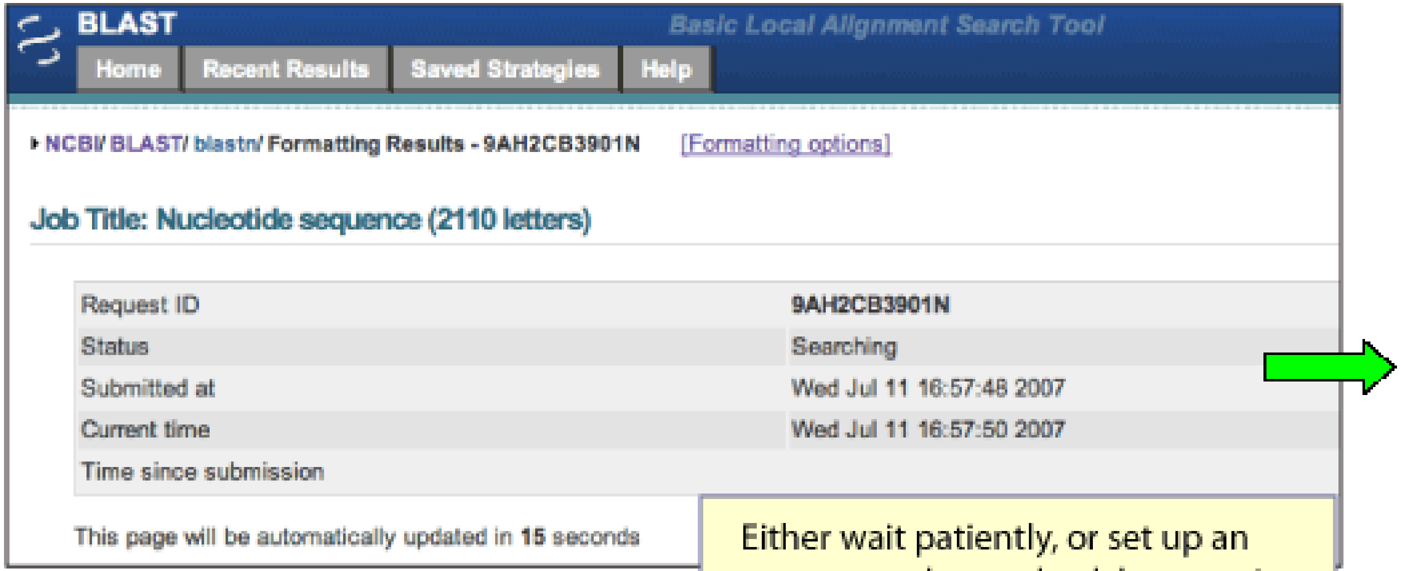
□ Show results in a new window

▶ Algorithm parameters

For the BLAST activities:

1. Change the database to "Others"
2. Select "nucleotide collection" from the pull-down menu.
3. Change the program selection to optimize for more dissimilar sequences.

BLAST Tutorial → Slide 4



NCBI/ BLAST/ blastn/ Formatting Results - 9AH2CB3901N [Formatting options]

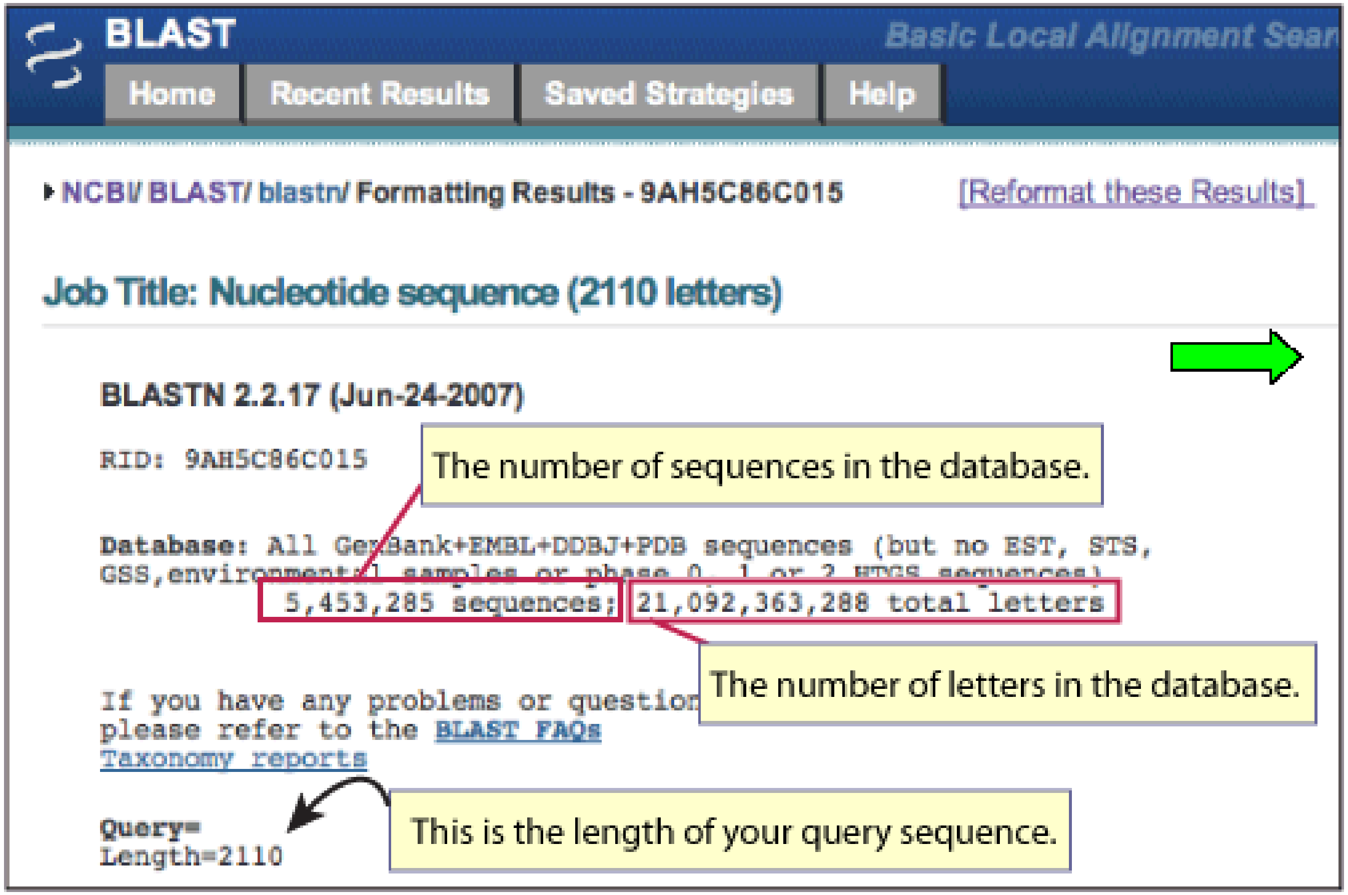
Job Title: Nucleotide sequence (2110 letters)

Request ID	9AH2CB3901N
Status	Searching
Submitted at	Wed Jul 11 16:57:48 2007
Current time	Wed Jul 11 16:57:50 2007
Time since submission	

This page will be automatically updated in **15** seconds

Either wait patiently, or set up an account and come back later to view your results.

BLAST Tutorial → Slide 5



NCBI/ BLAST/ blastn/ Formatting Results - 9AH5C86C015

Job Title: Nucleotide sequence (2110 letters)

BLASTN 2.2.17 (Jun-24-2007)

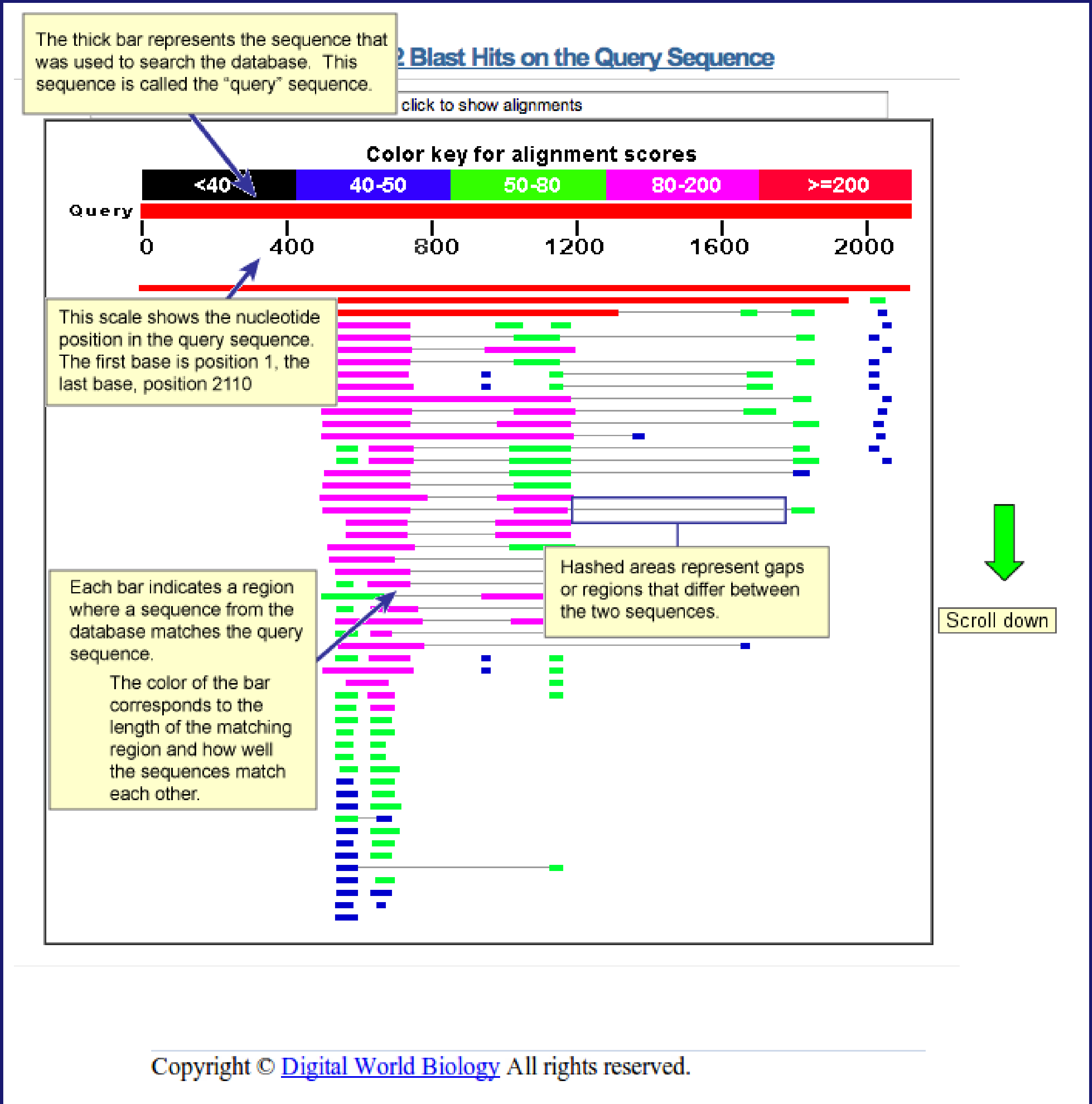
```
RID: 9AH5C86C015

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
           5,453,285 sequences; 21,092,363,288 total letters


If you have any problems or questions
please refer to the BLAST FAQs
Taxonomy reports

Query=
Length=2110
```

The number of sequences in the database.

The number of letters in the database.

This is the length of your query sequence.

BLAST Tutorial → Slide 6



The thick bar represents the sequence that was used to search the database. This sequence is called the "query" sequence.

This scale shows the nucleotide position in the query sequence. The first base is position 1, the last base, position 2110

Each bar indicates a region where a sequence from the database matches the query sequence.

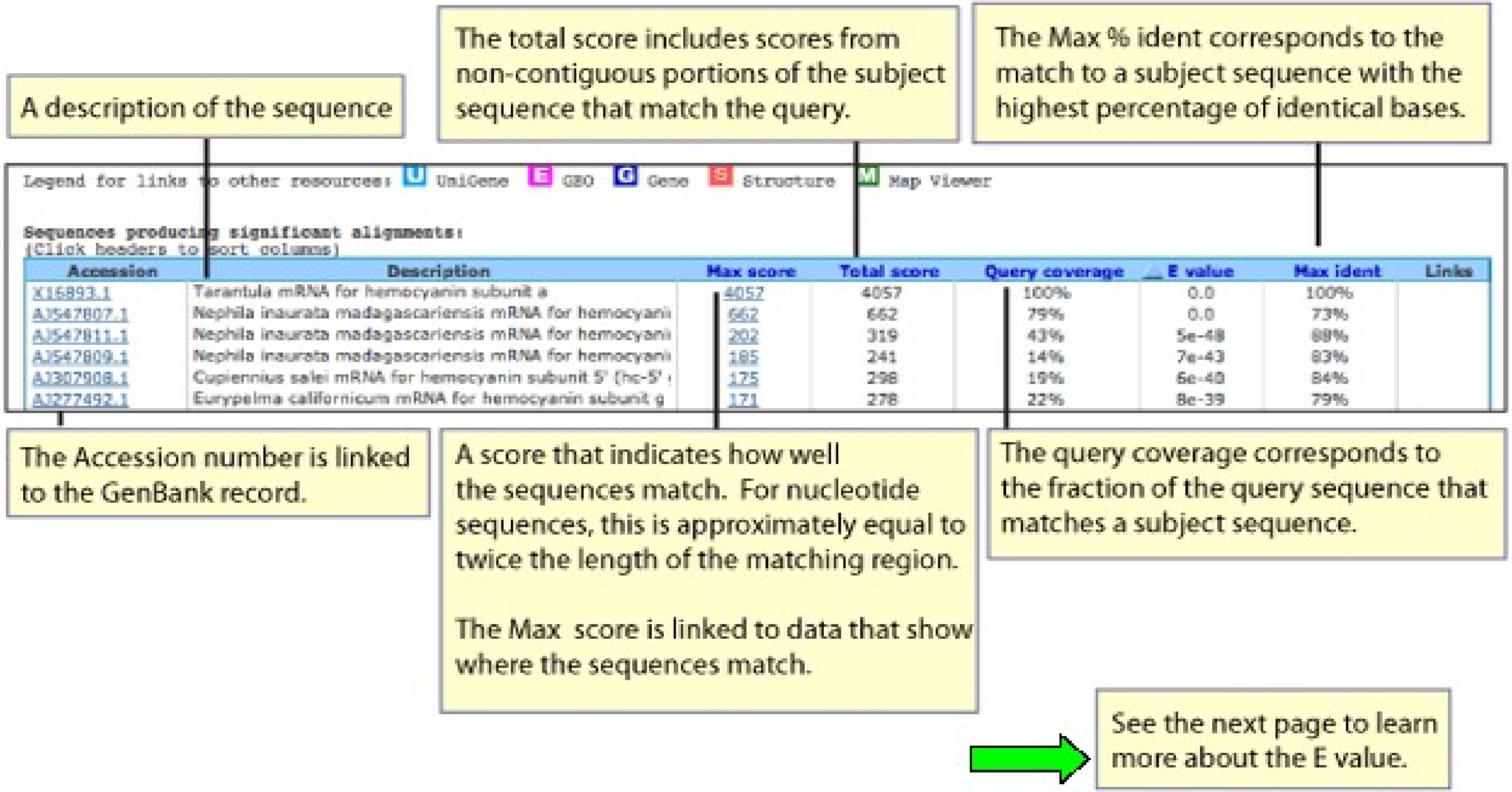
The color of the bar corresponds to the length of the matching region and how well the sequences match each other.

Hashed areas represent gaps or regions that differ between the two sequences.

Scroll down

Copyright © Digital World Biology All rights reserved.

BLAST Tutorial → Slide 7



A description of the sequence

The total score includes scores from non-contiguous portions of the subject sequence that match the query.

The Max % ident corresponds to the match to a subject sequence with the highest percentage of identical bases.

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
X16893.1	Tarantula mRNA for hemocyanin subunit a	4057	4057	100%	0.0	100%	
AJ547807.1	Nephila inaurata madagascariensis mRNA for hemocyani	662	662	79%	0.0	73%	
AJ547811.1	Nephila inaurata madagascariensis mRNA for hemocyani	202	319	43%	5e-48	88%	
AJ547809.1	Nephila inaurata madagascariensis mRNA for hemocyani	185	241	14%	7e-43	83%	
AJ307908.1	Cupiennius salei mRNA for hemocyanin subunit 5' (hc-5'	175	298	19%	6e-40	84%	
AJ277492.1	Eurypelma californicum mRNA for hemocyanin subunit g	171	278	22%	8e-39	79%	

The Accession number is linked to the GenBank record.

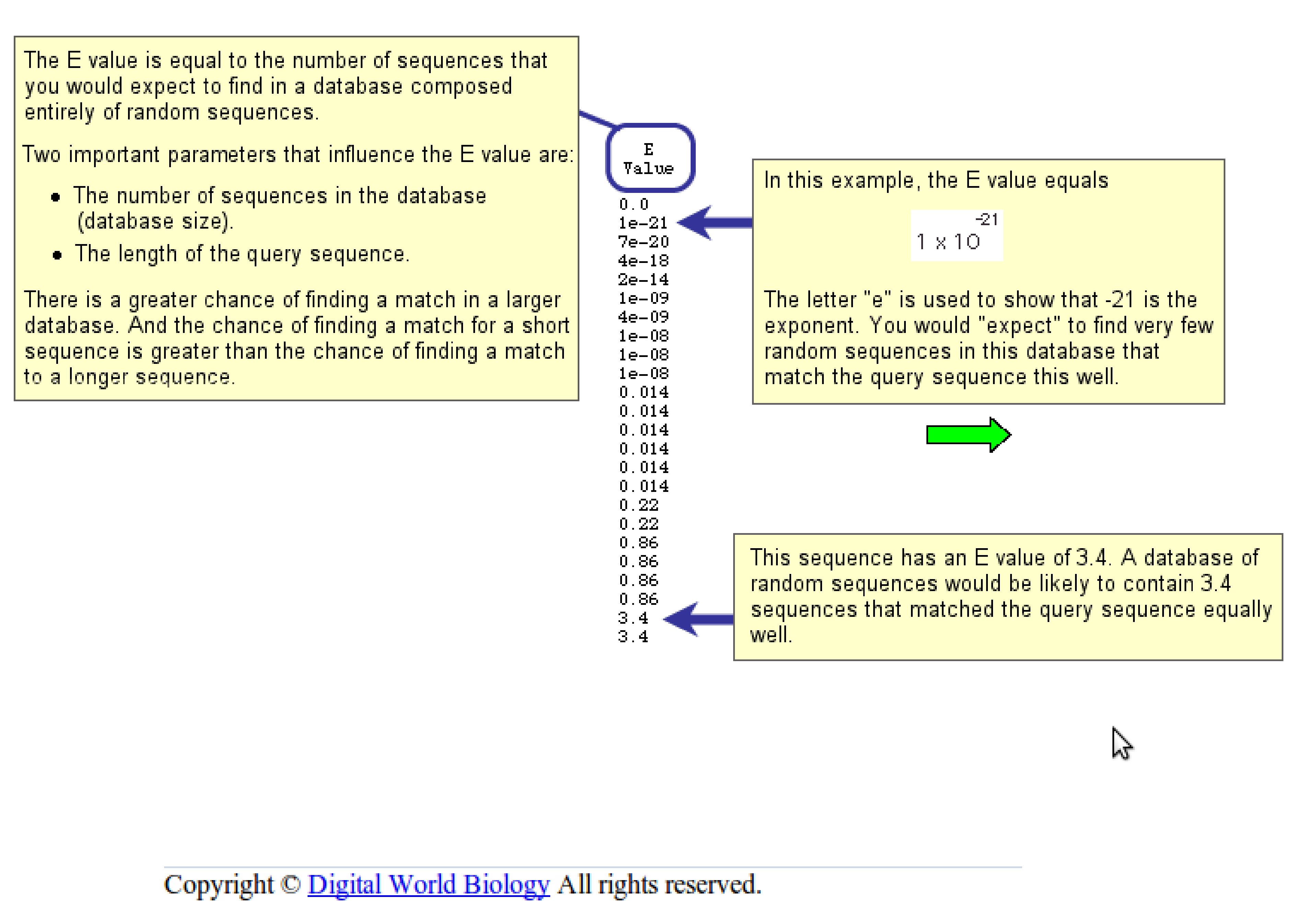
A score that indicates how well the sequences match. For nucleotide sequences, this is approximately equal to twice the length of the matching region.

The Max score is linked to data that show where the sequences match.

The query coverage corresponds to the fraction of the query sequence that matches a subject sequence.

See the next page to learn more about the E value.

BLAST Tutorial → Slide 8



The E value is equal to the number of sequences that you would expect to find in a database composed entirely of random sequences.

Two important parameters that influence the E value are:

- The number of sequences in the database (database size).
- The length of the query sequence.

There is a greater chance of finding a match in a larger database. And the chance of finding a match for a short sequence is greater than the chance of finding a match to a longer sequence.

```
  E
Value
0.0
1e-21
7e-20
4e-18
2e-14
1e-09
4e-09
1e-08
1e-08
1e-08
0.014
0.014
0.014
0.014
0.014
0.014
0.22
0.22
0.86
0.86
0.86
0.86
3.4
3.4
```

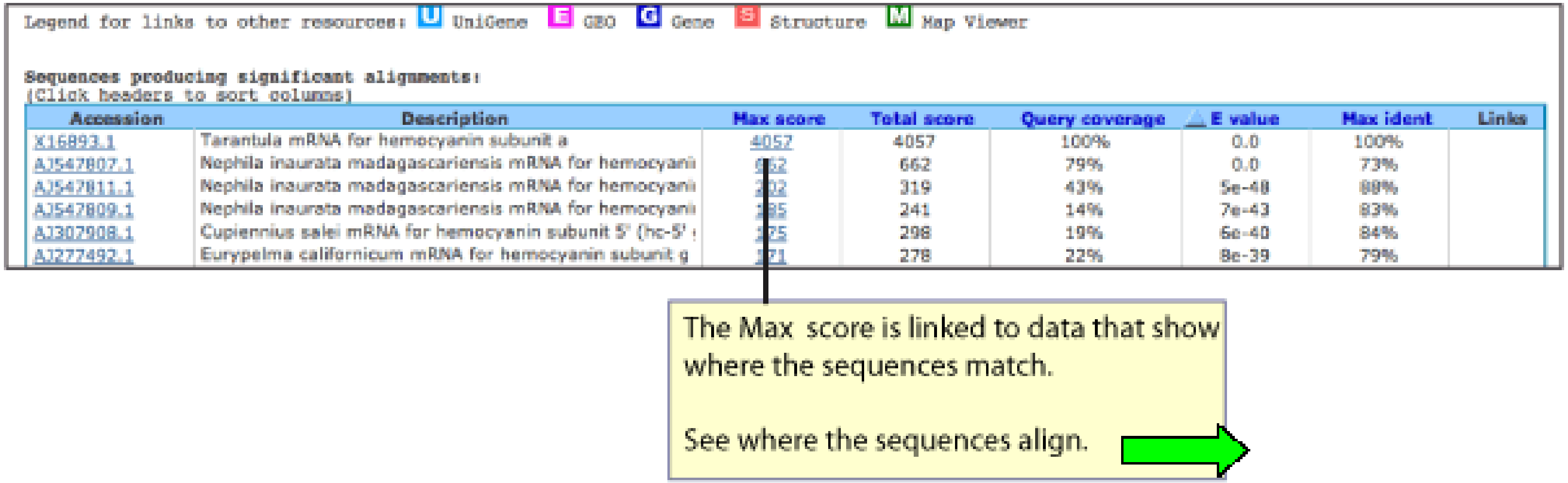
In this example, the E value equals

1×10^{-21}

The letter "e" is used to show that -21 is the exponent. You would "expect" to find very few random sequences in this database that match the query sequence this well.

This sequence has an E value of 3.4. A database of random sequences would be likely to contain 3.4 sequences that matched the query sequence equally well.

BLAST Tutorial → Slide 9



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

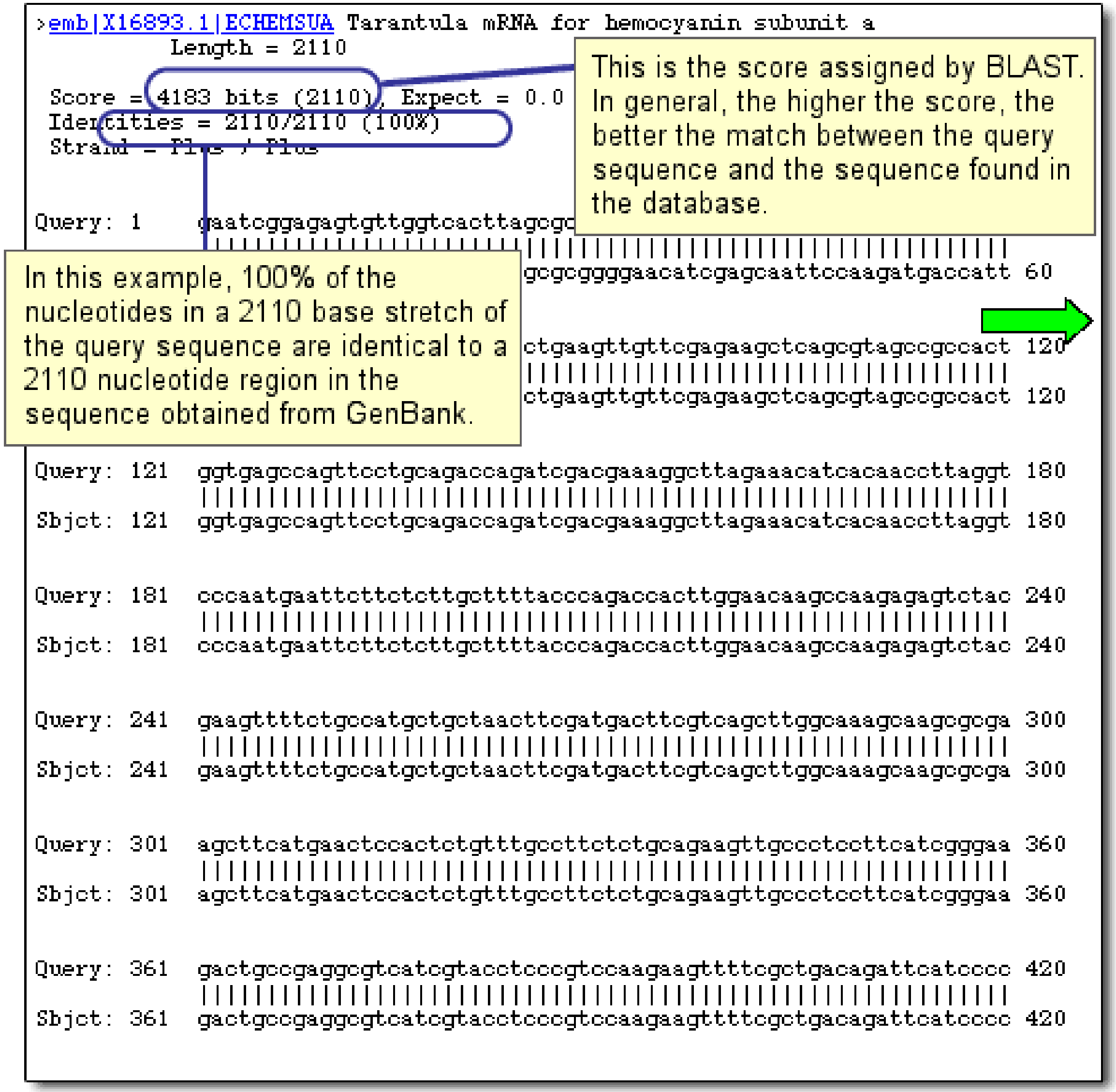
Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	△ E value	Max ident	Links
X16893.1	Tarantula mRNA for hemocyanin subunit a	4057	4057	100%	0.0	100%	
AJ547807.1	Nephila inaurata madagascariensis mRNA for hemocyani	662	662	79%	0.0	73%	
AJ547811.1	Nephila inaurata madagascariensis mRNA for hemocyani	202	319	43%	5e-48	88%	
AJ547809.1	Nephila inaurata madagascariensis mRNA for hemocyani	185	241	14%	7e-43	83%	
AJ307908.1	Cupiennius salei mRNA for hemocyanin subunit 5' (hc-5' ;	175	298	19%	6e-40	84%	
AJ277492.1	Eurypelma californicum mRNA for hemocyanin subunit g	171	278	22%	8e-39	79%	

The Max score is linked to data that show where the sequences match.

See where the sequences align.

BLAST Tutorial → Slide 10

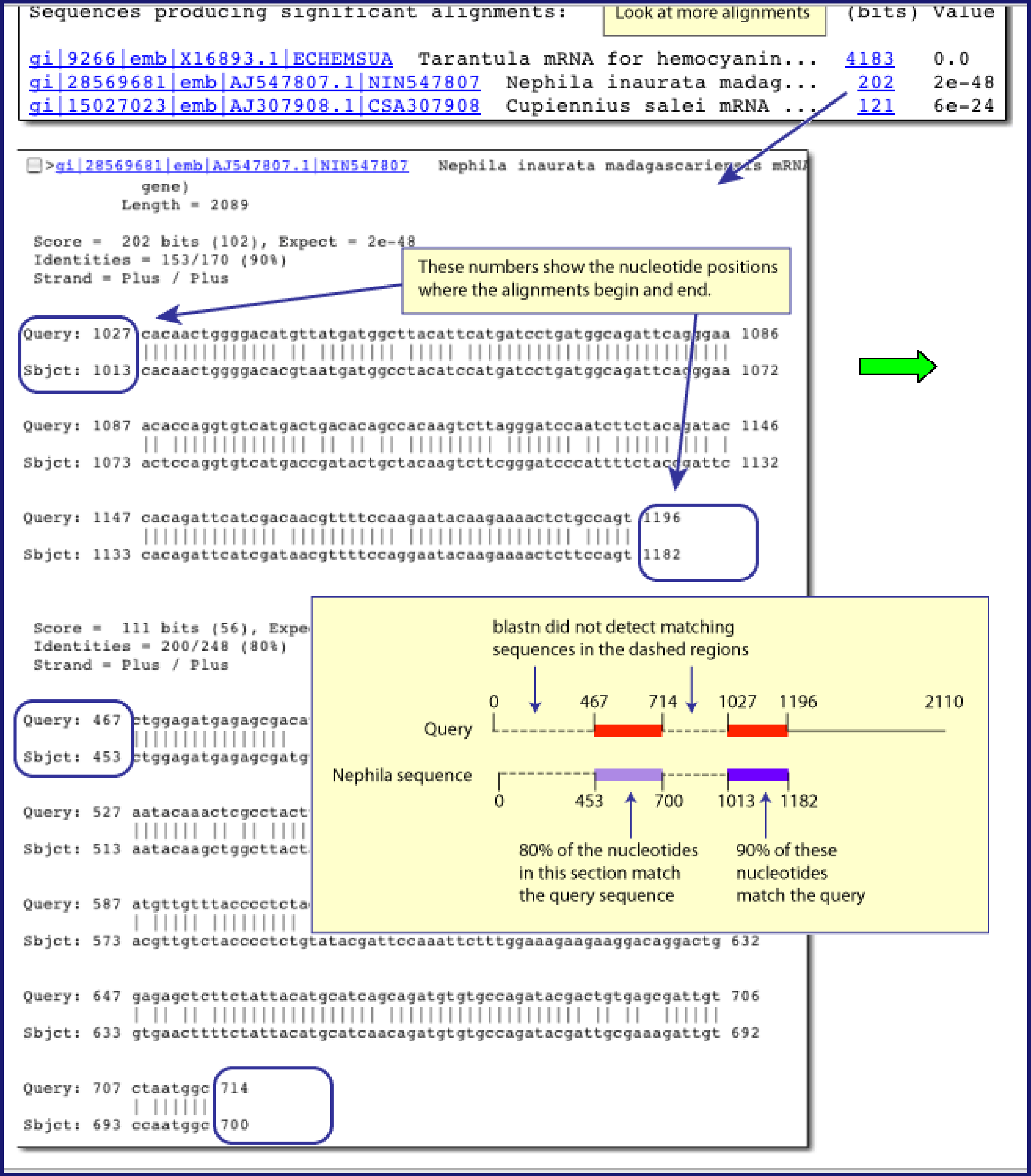


```
>emb|X16893.1|ECHEMSUA Tarantula mRNA for hemocyanin subunit a
          Length = 2110

 Score = 4183 bits (2110), Expect = 0.0
 Identities = 2110/2110 (100%)
 Strand = Plus / Plus


Query: 1     gaatcggagagtgttggtcacttagcgc
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
                                        gcgcggggaacatcgagcaattccaagatgaccatt 60


                                        ctgaagttgttcgagaagctcagcgtagccgccact 120
                                        ||||||||||||||||||||||||||||||||||||
                                        ctgaagttgttcgagaagctcagcgtagccgccact 120


Query: 121   ggtgagccagttcctgcagaccagatcgacgaaaggcttagaaacatcacaaccttaggt 180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 121   ggtgagccagttcctgcagaccagatcgacgaaaggcttagaaacatcacaaccttaggt 180


Query: 181   cccaatgaattcttctcttgcttttacccagaccacttggaacaagccaagagagtctac 240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 181   cccaatgaattcttctcttgcttttacccagaccacttggaacaagccaagagagtctac 240


Query: 241   gaagttttctgccatgctgctaacttcgatgacttcgtcagcttggcaaagcaagcgcga 300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 241   gaagttttctgccatgctgctaacttcgatgacttcgtcagcttggcaaagcaagcgcga 300


Query: 301   agcttcatgaactccactctgtttgccttctctgcagaagttgccctccttcatcgggaa 360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 301   agcttcatgaactccactctgtttgccttctctgcagaagttgccctccttcatcgggaa 360


Query: 361   gactgccgaggcgtcatcgtacctcccgtccaagaagttttcgctgacagattcatcccc 420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 361   gactgccgaggcgtcatcgtacctcccgtccaagaagttttcgctgacagattcatcccc 420
```

This is the 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.

In this example, 100% of the nucleotides in a 2110 base stretch of the query sequence are identical to a 2110 nucleotide region in the sequence obtained from GenBank.

BLAST Tutorial → Slide 10a



```
Sequences producing significant alignments:                    (bits) Value

gi|9266|emb|X16893.1|ECHEMSUA  Tarantula mRNA for hemocyanin...  4183    0.0
gi|28569681|emb|AJ547807.1|NIN547807  Nephila inaurata madag...   202    2e-48
gi|15027023|emb|AJ307908.1|CSA307908  Cupiennius salei mRNA ...   121    6e-24
```

Look at more alignments

```
>gi|28569681|emb|AJ547807.1|NIN547807   Nephila inaurata madagascariensis mRNA
          gene)
         Length = 2089

 Score =  202 bits (102), Expect = 2e-48
 Identities = 153/170 (90%)
 Strand = Plus / Plus

Query: 1027 cacaactggggacatgttatgatggcttacattcatgatcctgatggcagattcaggaa 1086
             |||||||||||||| || |||||||| ||||| |||||||||||||||||||||||||||
Sbjct: 1013 cacaactggggacacgtaatgatggcctacatccatgatcctgatggcagattcaggaa 1072

Query: 1087 acaccaggtgtcatgactgacacagccacaagtcttagggatccaatcttctacagatac 1146
             || |||||||||||||| || || || ||||||||| ||||||| || |||||| ||| |
Sbjct: 1073 actccaggtgtcatgaccgatactgctacaagtcttcgggatcccattttctacggattc 1132

Query: 1147 cacagattcatcgacaacgttttccaagaatacaagaaaactctgccagt 1196
             |||||||||||||| ||||||||||| ||||||||||||||||| |||||
Sbjct: 1133 cacagattcatcgataacgttttccaggaatacaagaaaactcttccagt 1182


 Score =  111 bits (56), Expe
 Identities = 200/248 (80%)
 Strand = Plus / Plus

Query: 467 ctggagatgagagcgaca
            ||||||||||||||||
Sbjct: 453 ctggagatgagagcgatg

Query: 527 aatacaaactcgcctact
            ||||||| || || ||||
Sbjct: 513 aatacaagctggcttact

Query: 587 atgttgtttacccctcta
            | ||||| |||||||||
Sbjct: 573 acgttgtctacccctctgtatacgattccaaattctttggaaagaagaaggacaggactg 632

Query: 647 gagagctcttctattacatgcatcagcagatgtgtgccagatacgactgtgagcgattgt 706
            | || || ||||||||||||||||| |||||||||||||||||||| || ||  ||||||
Sbjct: 633 gtgaacttttctattacatgcatcaacagatgtgtgccagatacgattgcgaaagattgt 692

Query: 707 ctaatggc 714
            | ||||||
Sbjct: 693 ccaatggc 700
```

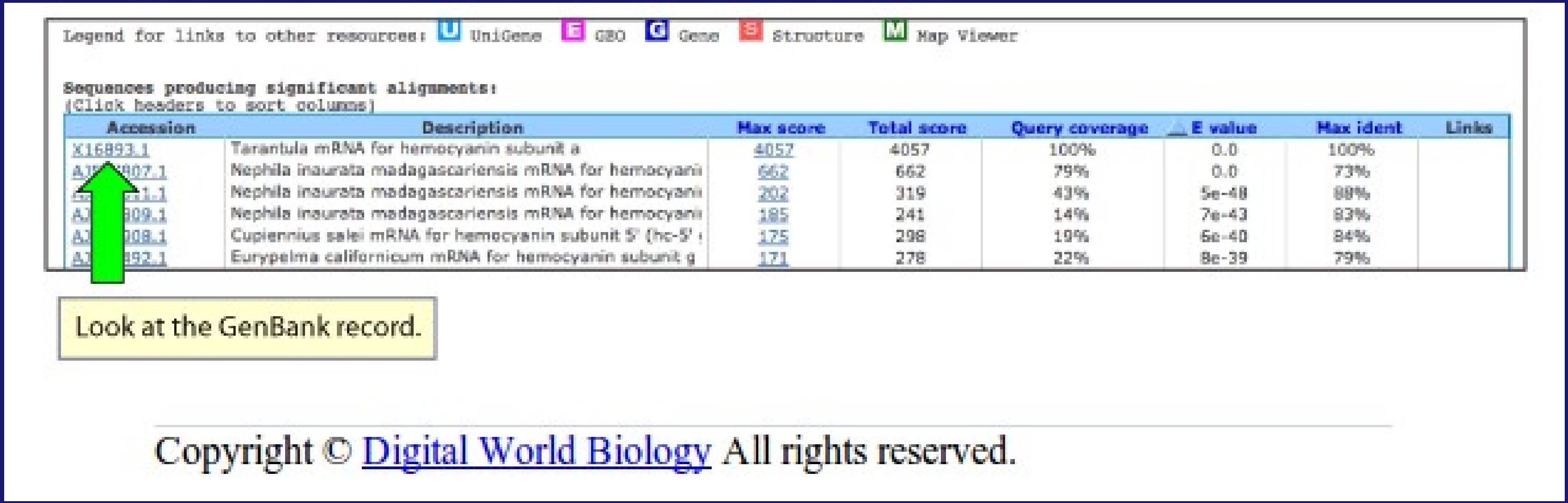
These numbers show the nucleotide positions where the alignments begin and end.

blastn did not detect matching sequences in the dashed regions

80% of the nucleotides in this section match the query sequence

90% of these nucleotides match the query

BLAST Tutorial → Slide 10b



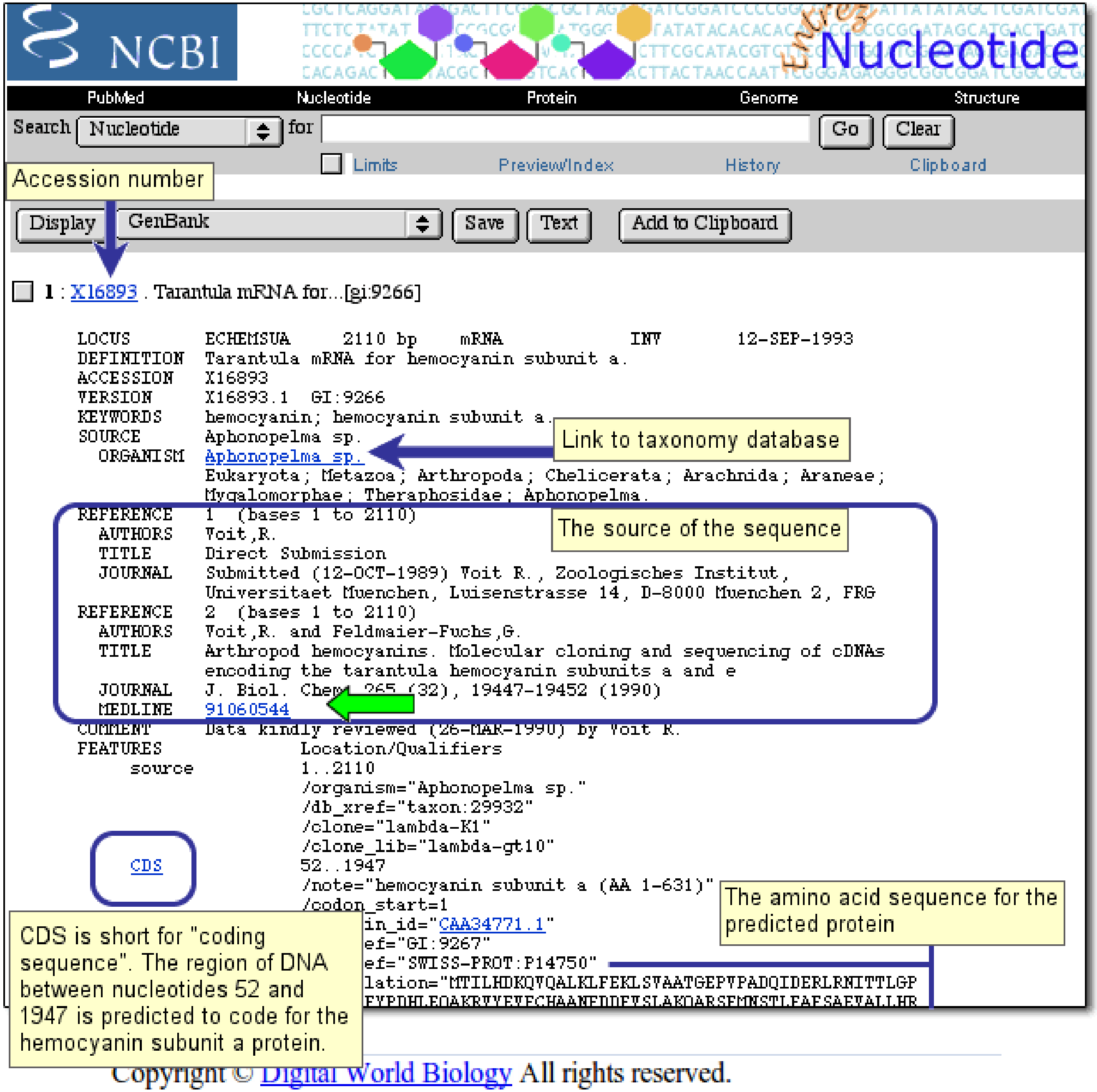
Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

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

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
X16893.1	Tarantula mRNA for hemocyanin subunit a	4057	4057	100%	0.0	100%	
AJ[illegible]807.1	Nephila inaurata madagascariensis mRNA for hemocyani	662	662	79%	0.0	73%	
A[illegible]1.1	Nephila inaurata madagascariensis mRNA for hemocyani	202	319	43%	5e-48	88%	
AJ[illegible]809.1	Nephila inaurata madagascariensis mRNA for hemocyani	185	241	14%	7e-43	83%	
AJ[illegible]808.1	Cupiennius salei mRNA for hemocyanin subunit 5' (hc-5' ;	175	298	19%	6e-40	84%	
AJ[illegible]892.1	Eurypelma californicum mRNA for hemocyanin subunit g	171	278	22%	8e-39	79%	

Look at the GenBank record.

BLAST Tutorial → Slide 11



Accession number

□ 1 : X16893 . Tarantula mRNA for...[gi:9266]

```
LOCUS       ECHEMSUA     2110 bp    mRNA            INV       12-SEP-1993
DEFINITION  Tarantula mRNA for hemocyanin subunit a.
ACCESSION   X16893
VERSION     X16893.1  GI:9266
KEYWORDS    hemocyanin; hemocyanin subunit a.
SOURCE      Aphonopelma sp.
  ORGANISM  Aphonopelma sp.
            Eukaryota; Metazoa; Arthropoda; Chelicerata; Arachnida; Araneae;
            Mygalomorphae; Theraphosidae; Aphonopelma.
REFERENCE   1  (bases 1 to 2110)
  AUTHORS   Voit,R.
  TITLE     Direct Submission
  JOURNAL   Submitted (12-OCT-1989) Voit R., Zoologisches Institut,
            Universitaet Muenchen, Luisenstrasse 14, D-8000 Muenchen 2, FRG
REFERENCE   2  (bases 1 to 2110)
  AUTHORS   Voit,R. and Feldmaier-Fuchs,G.
  TITLE     Arthropod hemocyanins. Molecular cloning and sequencing of cDNAs
            encoding the tarantula hemocyanin subunits a and e
  JOURNAL   J. Biol. Chem. 265 (32), 19447-19452 (1990)
  MEDLINE   91060544
COMMENT     Data kindly reviewed (26-MAR-1990) by Voit R.
FEATURES             Location/Qualifiers
     source          1..2110
                     /organism="Aphonopelma sp."
                     /db_xref="taxon:29932"
                     /clone="lambda-K1"
                     /clone_lib="lambda-gt10"
     CDS             52..1947
                     /note="hemocyanin subunit a (AA 1-631)"
                     /codon_start=1
                     in_id="CAA34771.1"
                     ef="GI:9267"
                     ef="SWISS-PROT:P14750"
                     lation="MTILHDKQVQALKLFEKLSVAATGEPVPADQIDERLRNITTLGP
                     FYPDHLEQAKRVYEVFCHAANFDDFVSLAKQARSFMNSTLFAFSAFVALLHR
```

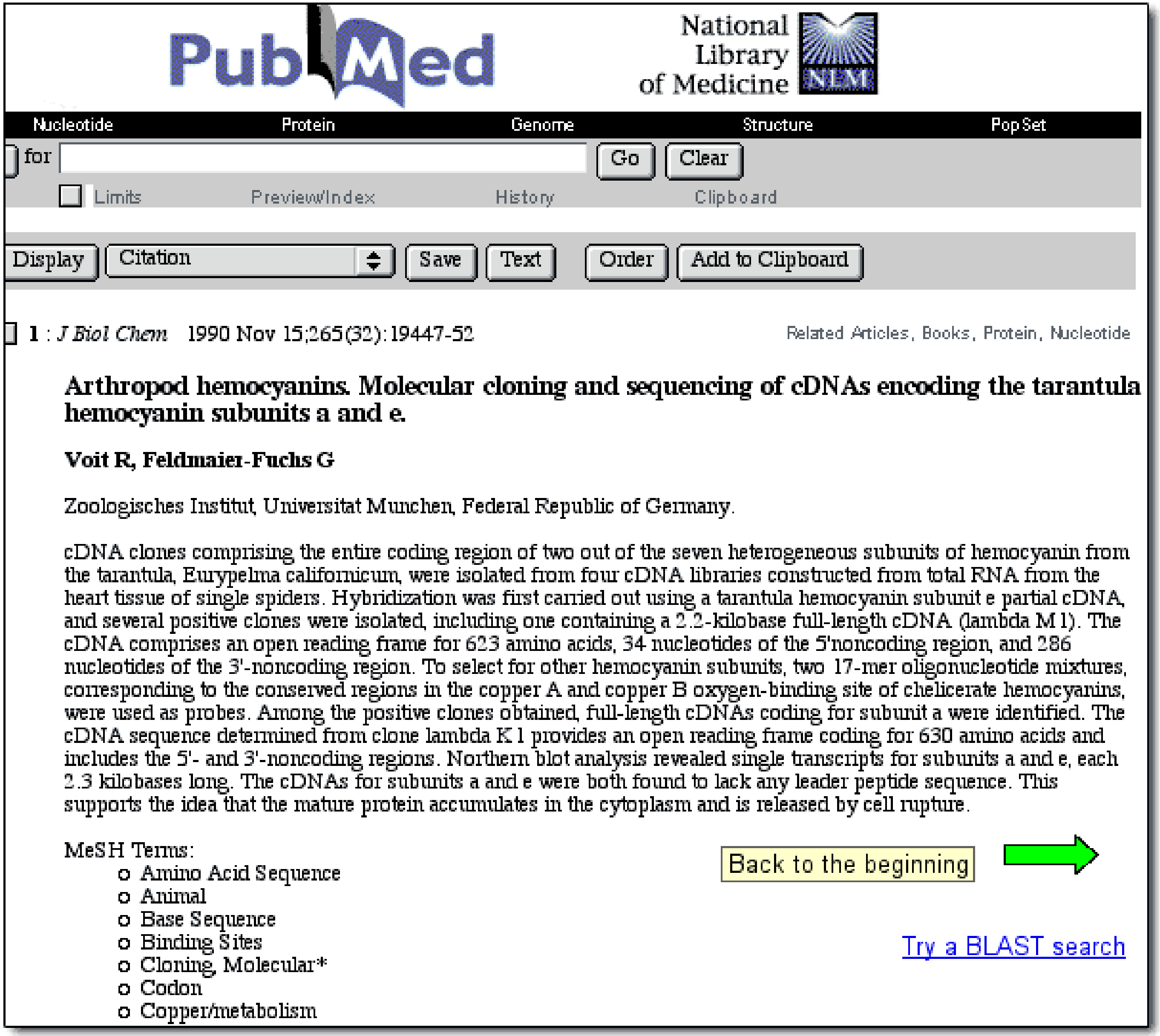
Link to taxonomy database

The source of the sequence

The amino acid sequence for the predicted protein

CDS is short for "coding sequence". The region of DNA between nucleotides 52 and 1947 is predicted to code for the hemocyanin subunit a protein.

BLAST Tutorial → Slide 12



1 : *J Biol Chem* 1990 Nov 15;265(32):19447-52

Arthropod hemocyanins. Molecular cloning and sequencing of cDNAs encoding the tarantula hemocyanin subunits a and e.

Voit R, Feldmaier-Fuchs G

Zoologisches Institut, Universitat Munchen, Federal Republic of Germany.

cDNA clones comprising the entire coding region of two out of the seven heterogeneous subunits of hemocyanin from the tarantula, Eurypelma californicum, were isolated from four cDNA libraries constructed from total RNA from the heart tissue of single spiders. Hybridization was first carried out using a tarantula hemocyanin subunit e partial cDNA, and several positive clones were isolated, including one containing a 2.2-kilobase full-length cDNA (lambda M 1). The cDNA comprises an open reading frame for 623 amino acids, 34 nucleotides of the 5'noncoding region, and 286 nucleotides of the 3'-noncoding region. To select for other hemocyanin subunits, two 17-mer oligonucleotide mixtures, corresponding to the conserved regions in the copper A and copper B oxygen-binding site of chelicerate hemocyanins, were used as probes. Among the positive clones obtained, full-length cDNAs coding for subunit a were identified. The cDNA sequence determined from clone lambda K 1 provides an open reading frame coding for 630 amino acids and includes the 5'- and 3'-noncoding regions. Northern blot analysis revealed single transcripts for subunits a and e, each 2.3 kilobases long. The cDNAs for subunits a and e were both found to lack any leader peptide sequence. This supports the idea that the mature protein accumulates in the cytoplasm and is released by cell rupture.

MeSH Terms:

- Amino Acid Sequence
- Animal
- Base Sequence
- Binding Sites
- Cloning, Molecular*
- Codon
- Copper/metabolism

Back to the beginning

Try a BLAST search

Hand-on BLAST Tutorial

(1) Open the BLAST web application:

http://blast.ncbi.nlm.nih.gov/Blast.cgi

(2) In another tab, open this web page:

http://www.digitalworldbiology.com/BLAST/62000sequences.html

(3) Copy and paste the example sequence into the text box on the BLAST page

Questions?

Questions?