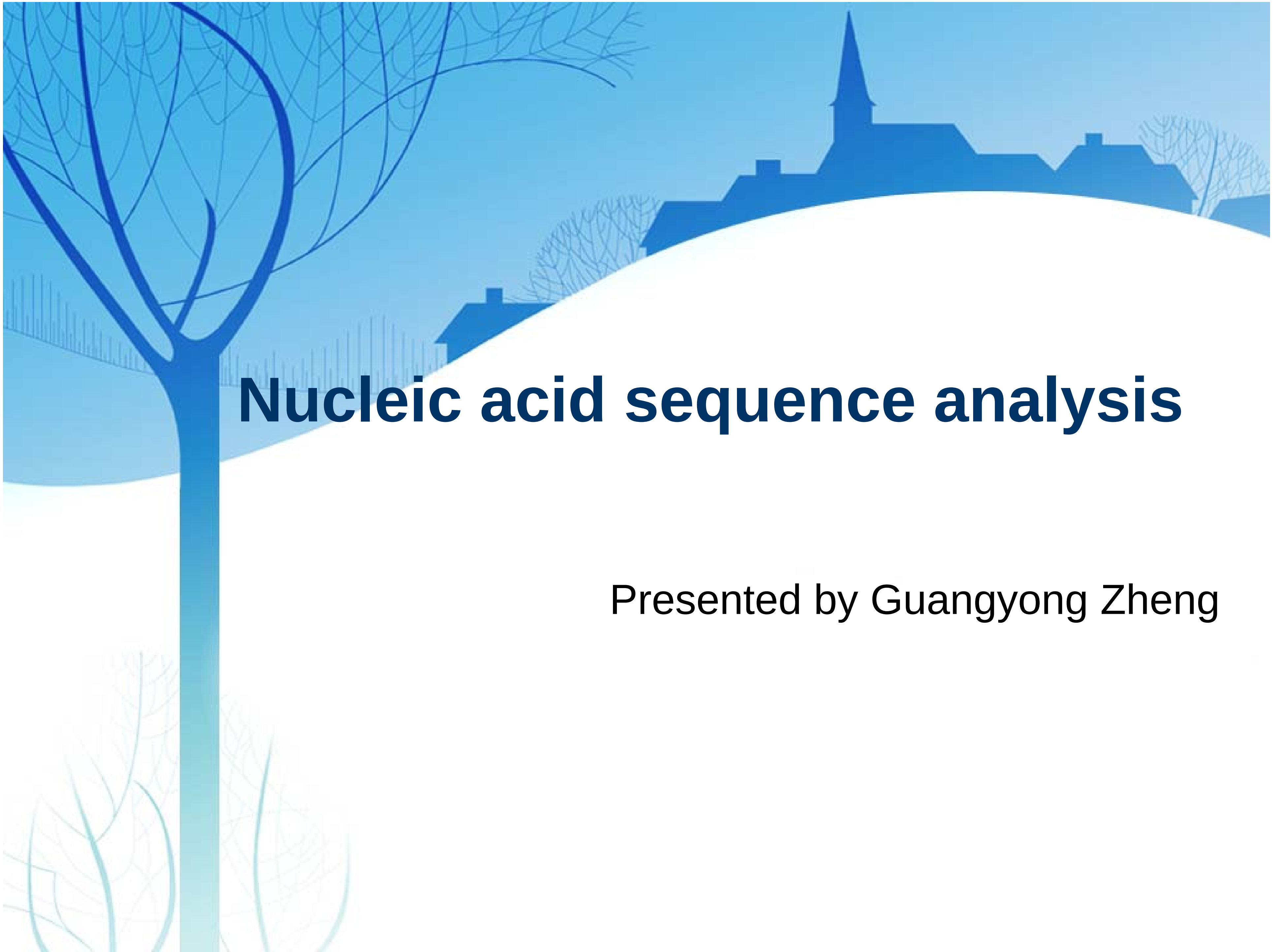


Nucleic acid sequence analysis

Presented by Guangyong Zheng

Outline

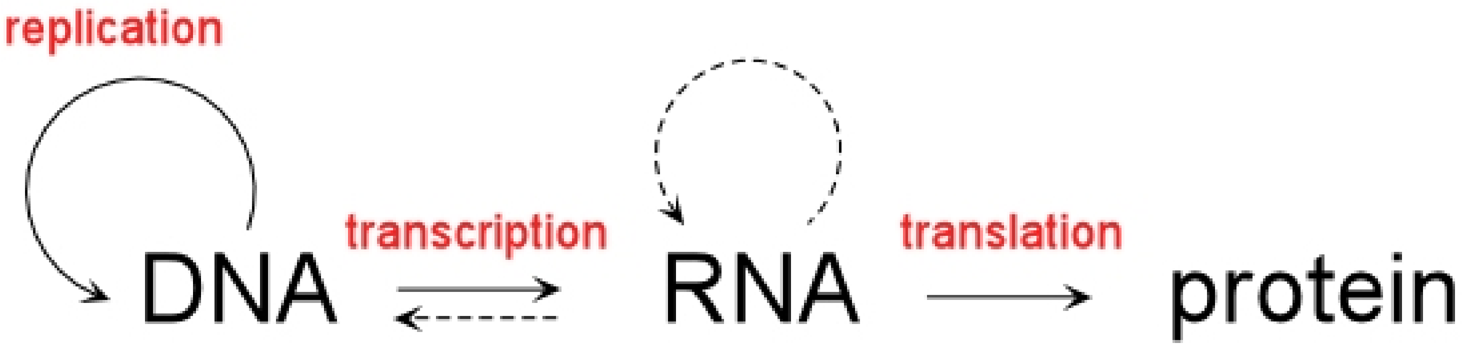
Nucleic acid sequence analysis

- Famous database of nucleic acid sequence
- Analysis of nucleic acid sequence
- BLAST software
  - Basic concept of BLAST
  - BLAST use - web manners
  - BLAST use – local manners

PART I Famous database of nucleic acid sequence

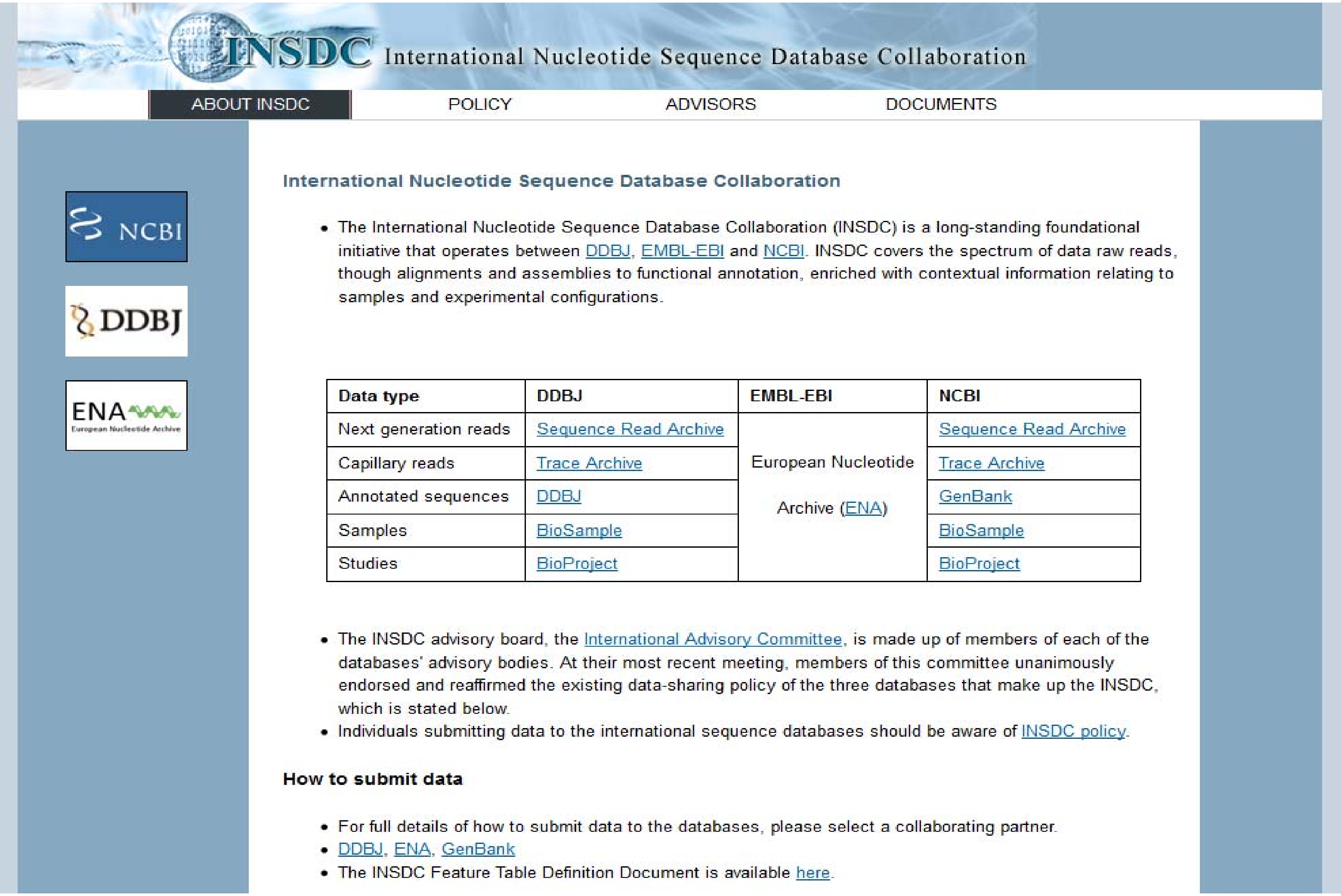
Database of nucleic acid sequence

- Nucleic acid sequence is important material of bioinformatics research.
- Nucleic acid sequences provide the fundamental starting point for describing and understanding the structure, function, and development of genetically diverse organisms.



Database of nucleic acid sequence

http://www.insdc.org/


INSDC International Nucleotide Sequence Database Collaboration
ABOUT INSDC
POLICY
ADVISORS
DOCUMENTS
NCBI
DDBJ
ENA
European Nucleotide Archive

International Nucleotide Sequence Database Collaboration

- The International Nucleotide Sequence Database Collaboration (INSDC) is a long-standing foundational initiative that operates between DDBJ, EMBL-EBI and NCBI. INSDC covers the spectrum of data raw reads, though alignments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations.

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

- The INSDC advisory board, the International Advisory Committee, is made up of members of each of the databases' advisory bodies. At their most recent meeting, members of this committee unanimously endorsed and reaffirmed the existing data-sharing policy of the three databases that make up the INSDC, which is stated below.
- Individuals submitting data to the international sequence databases should be aware of INSDC policy.

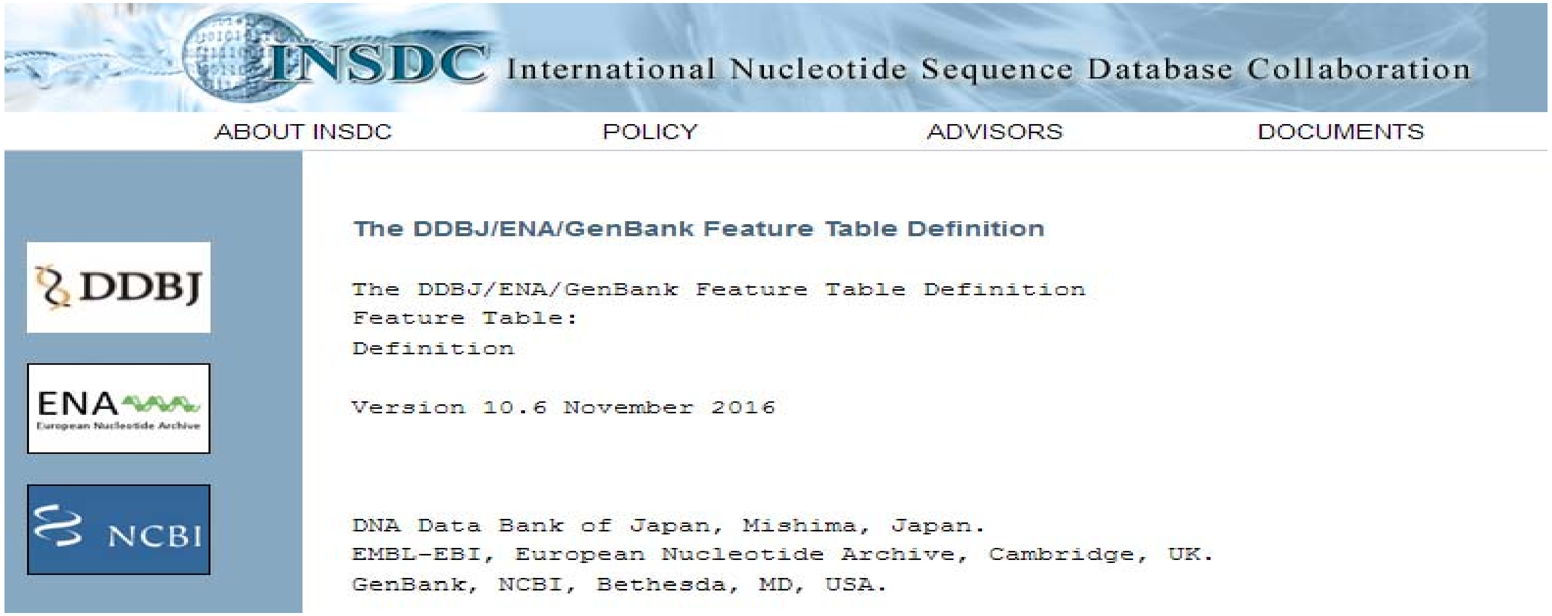
How to submit data

- For full details of how to submit data to the databases, please select a collaborating partner.
- DDBJ, ENA, GenBank
- The INSDC Feature Table Definition Document is available here.

Database of nucleic acid sequence

Feature table

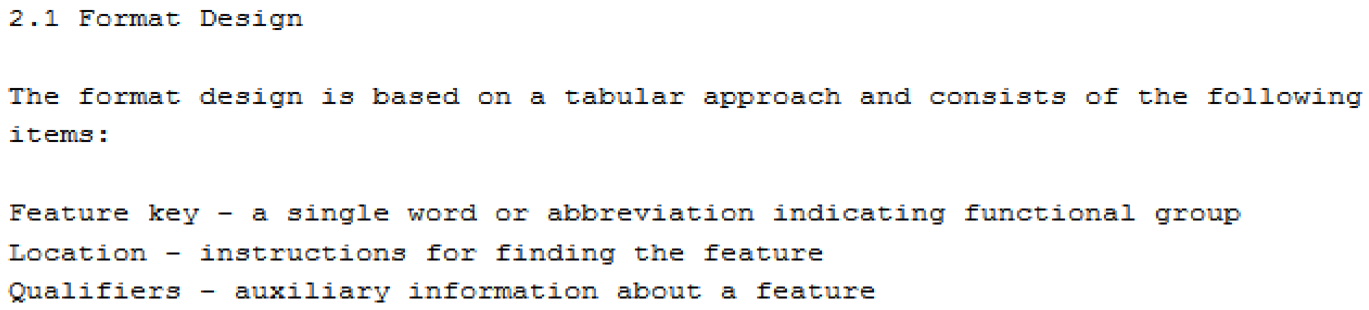
http://www.insdc.org/documents/feature_table.html



INSDC International Nucleotide Sequence Database Collaboration
ABOUT INSDC
POLICY
ADVISORS
DOCUMENTS
DDBJ
ENA European Nucleotide Archive
NCBI

The DDBJ/ENA/GenBank Feature Table Definition

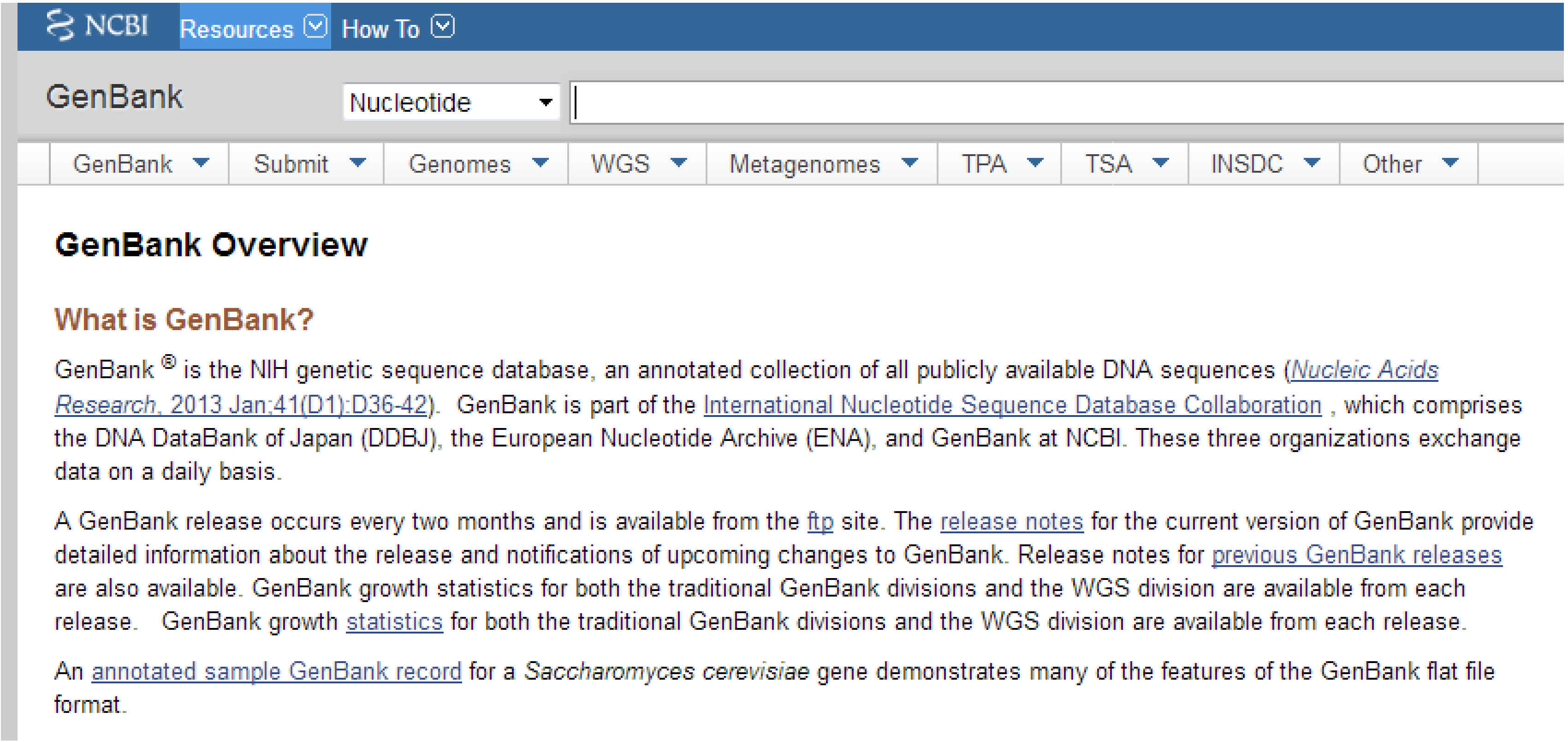
```
The DDBJ/ENA/GenBank Feature Table Definition
Feature Table:
Definition

Version 10.6 November 2016



DNA Data Bank of Japan, Mishima, Japan.
EMBL-EBI, European Nucleotide Archive, Cambridge, UK.
GenBank, NCBI, Bethesda, MD, USA.
```



```
2.1 Format Design

The format design is based on a tabular approach and consists of the following
items:

Feature key - a single word or abbreviation indicating functional group
Location - instructions for finding the feature
Qualifiers - auxiliary information about a feature
```

Database of nucleic acid sequence

http://www.ncbi.nlm.nih.gov/genbank/



GenBank Overview

What is GenBank?

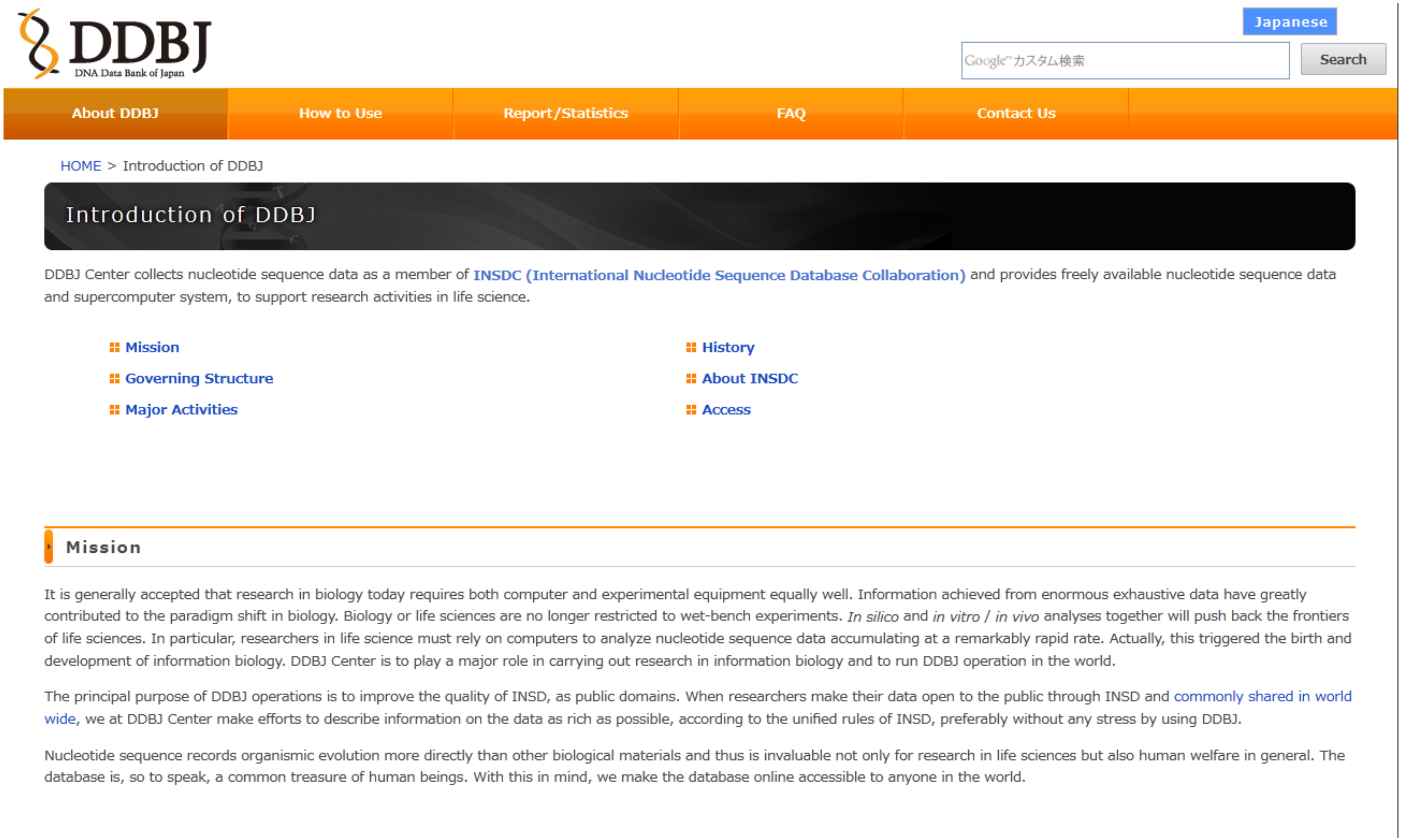
GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2013 Jan;41(D1):D36-42). GenBank is part of the International Nucleotide Sequence Database Collaboration , which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the ftp site. The release notes for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for previous GenBank releases are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release.

An annotated sample GenBank record for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Database of nucleic acid sequence

http://www.ddbj.nig.ac.jp



Introduction of DDBJ

DDBJ Center collects nucleotide sequence data as a member of INSDC (International Nucleotide Sequence Database Collaboration) and provides freely available nucleotide sequence data and supercomputer system, to support research activities in life science.

Mission

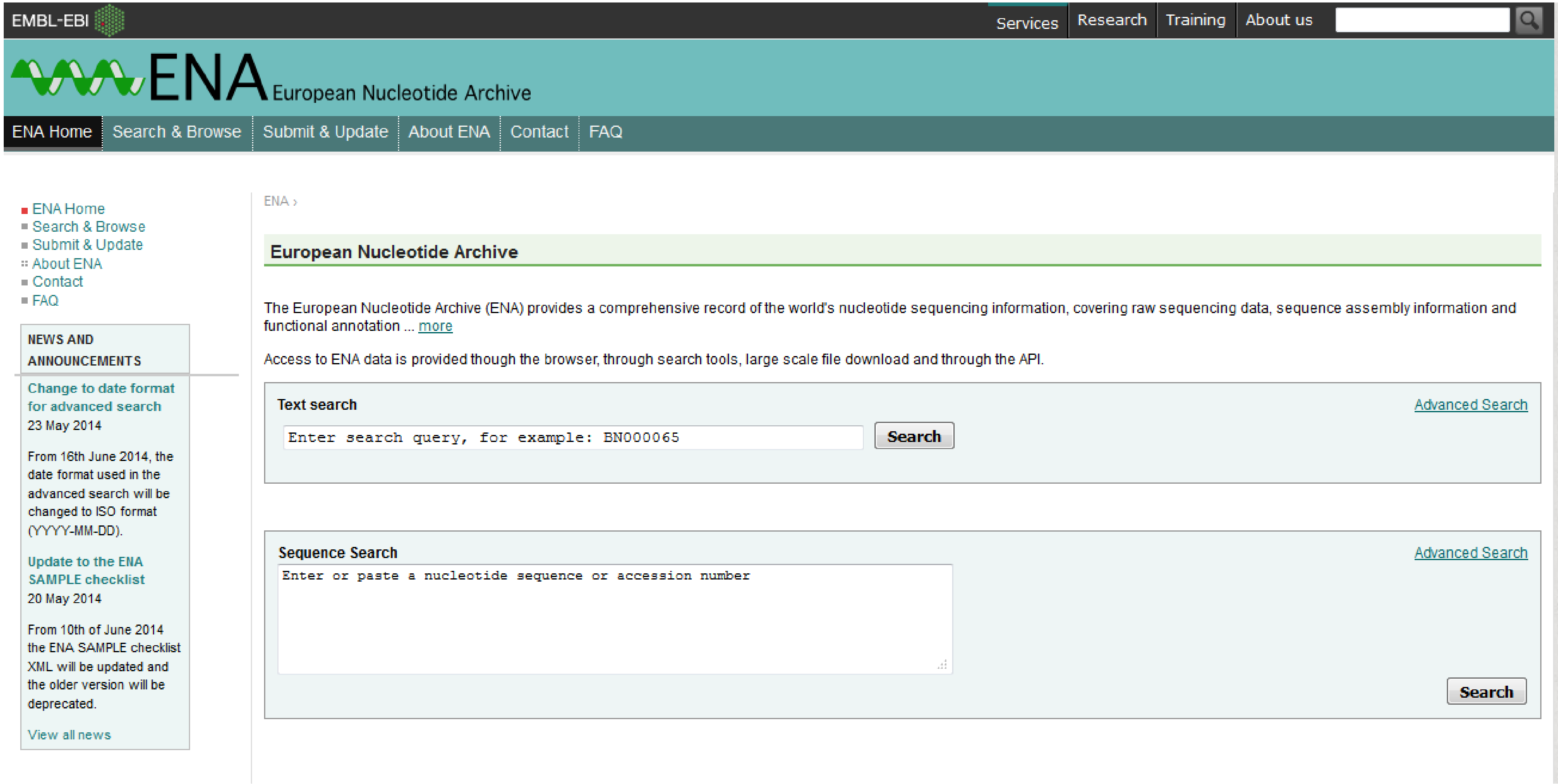
It is generally accepted that research in biology today requires both computer and experimental equipment equally well. Information achieved from enormous exhaustive data have greatly contributed to the paradigm shift in biology. Biology or life sciences are no longer restricted to wet-bench experiments. *In silico* and *in vitro / in vivo* analyses together will push back the frontiers of life sciences. In particular, researchers in life science must rely on computers to analyze nucleotide sequence data accumulating at a remarkably rapid rate. Actually, this triggered the birth and development of information biology. DDBJ Center is to play a major role in carrying out research in information biology and to run DDBJ operation in the world.

The principal purpose of DDBJ operations is to improve the quality of INSD, as public domains. When researchers make their data open to the public through INSD and commonly shared in world wide, we at DDBJ Center make efforts to describe information on the data as rich as possible, according to the unified rules of INSD, preferably without any stress by using DDBJ.

Nucleotide sequence records organismic evolution more directly than other biological materials and thus is invaluable not only for research in life sciences but also human welfare in general. The database is, so to speak, a common treasure of human beings. With this in mind, we make the database online accessible to anyone in the world.

Database of nucleic acid sequence

http://www.ebi.ac.uk/ena/



- ENA Home
- Search & Browse
- Submit & Update
- About ENA
- Contact
- FAQ

NEWS AND ANNOUNCEMENTS

Change to date format for advanced search
23 May 2014

From 16th June 2014, the date format used in the advanced search will be changed to ISO format (YYYY-MM-DD).

Update to the ENA SAMPLE checklist
20 May 2014

From 10th of June 2014 the ENA SAMPLE checklist XML will be updated and the older version will be deprecated.

View all news

ENA ›

European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation ... more

Access to ENA data is provided though the browser, through search tools, large scale file download and through the API.

Text search Advanced Search

Enter search query, for example: BN000065 [Search]

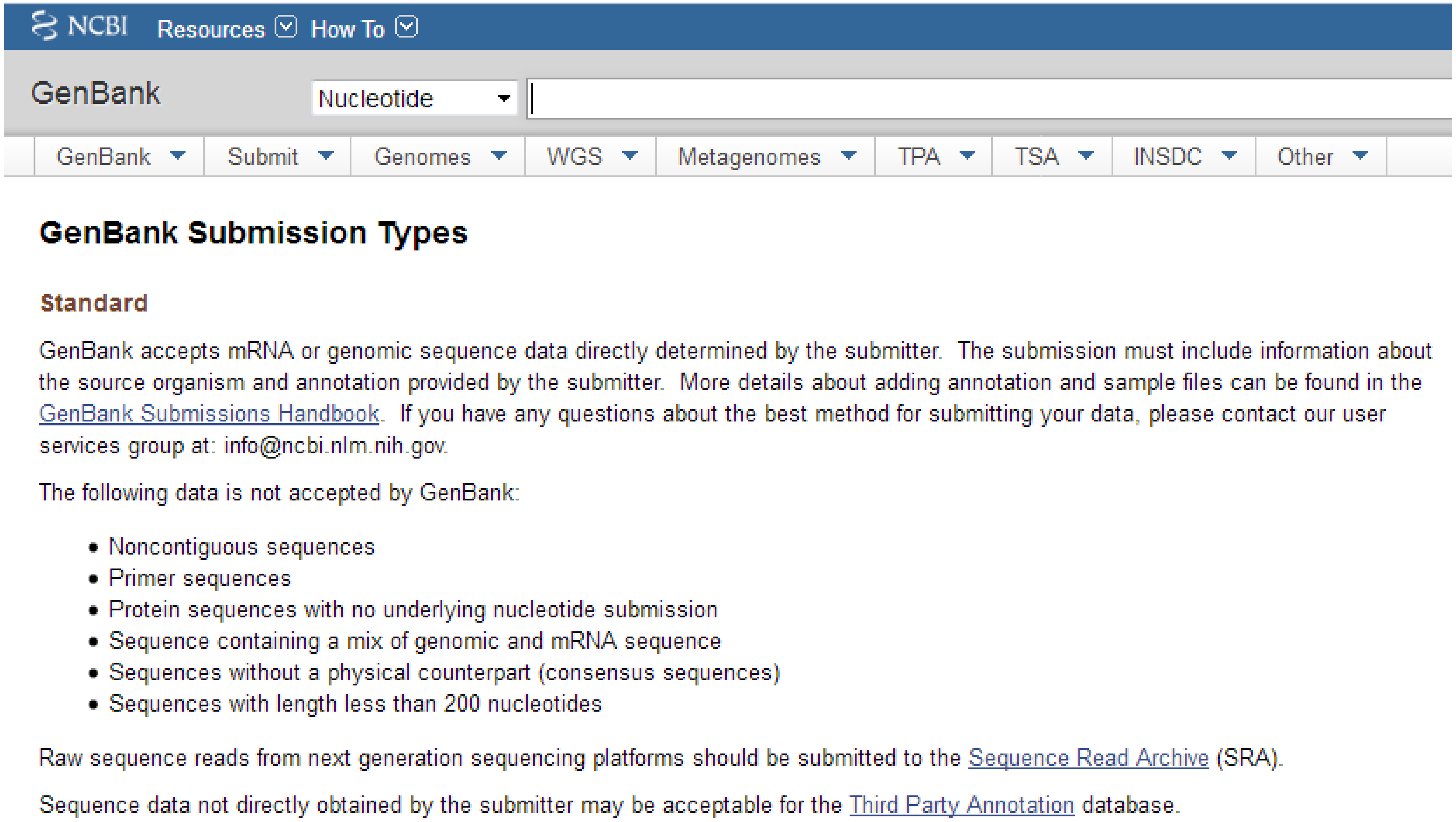
Sequence Search Advanced Search

Enter or paste a nucleotide sequence or accession number

[Search]

Database of nucleic acid sequence

Submit sequence (http://www.ncbi.nlm.nih.gov/genbank/submit_types)



GenBank Submission Types

Standard

GenBank accepts mRNA or genomic sequence data directly determined by the submitter. The submission must include information about the source organism and annotation provided by the submitter. More details about adding annotation and sample files can be found in the GenBank Submissions Handbook. If you have any questions about the best method for submitting your data, please contact our user services group at: info@ncbi.nlm.nih.gov.

The following data is not accepted by GenBank:

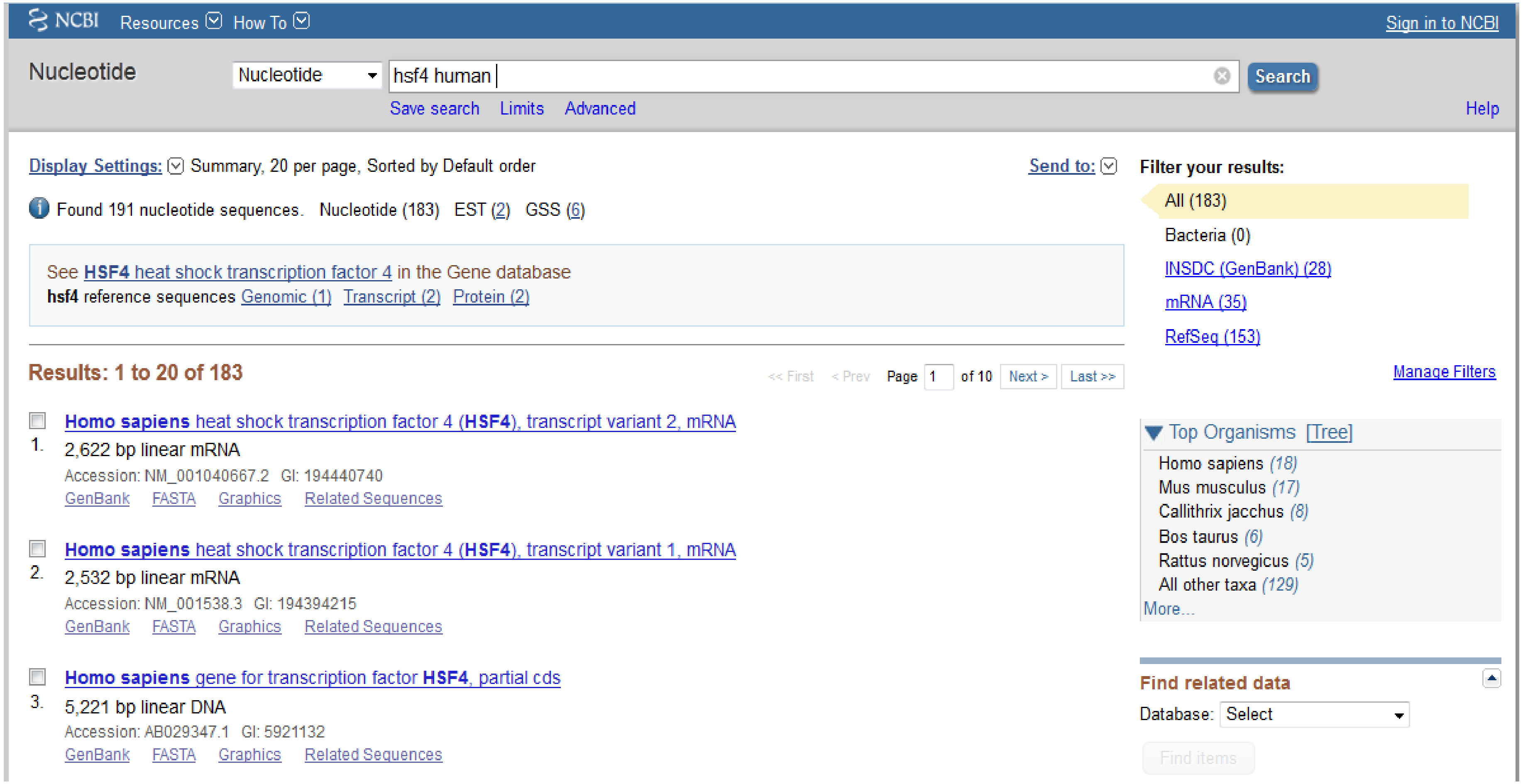
- Noncontiguous sequences
- Primer sequences
- Protein sequences with no underlying nucleotide submission
- Sequence containing a mix of genomic and mRNA sequence
- Sequences without a physical counterpart (consensus sequences)
- Sequences with length less than 200 nucleotides

Raw sequence reads from next generation sequencing platforms should be submitted to the Sequence Read Archive (SRA).

Sequence data not directly obtained by the submitter may be acceptable for the Third Party Annotation database.

Getting sequence from the GenBank database

http://www.ncbi.nlm.nih.gov/nucleotide/



hsf4 human

Display Settings: Summary, 20 per page, Sorted by Default order

Send to:

Found 191 nucleotide sequences. Nucleotide (183) EST (2) GSS (6)

See **HSF4** heat shock transcription factor 4 in the Gene database
hsf4 reference sequences Genomic (1) Transcript (2) Protein (2)

Results: 1 to 20 of 183

<< First < Prev Page 1 of 10 Next > Last >>

1. **Homo sapiens** heat shock transcription factor 4 (**HSF4**), transcript variant 2, mRNA
   2,622 bp linear mRNA
   Accession: NM_001040667.2 GI: 194440740
   GenBank FASTA Graphics Related Sequences

2. **Homo sapiens** heat shock transcription factor 4 (**HSF4**), transcript variant 1, mRNA
   2,532 bp linear mRNA
   Accession: NM_001538.3 GI: 194394215
   GenBank FASTA Graphics Related Sequences

3. **Homo sapiens** gene for transcription factor **HSF4**, partial cds
   5,221 bp linear DNA
   Accession: AB029347.1 GI: 5921132
   GenBank FASTA Graphics Related Sequences

Filter your results:

All (183)
Bacteria (0)
INSDC (GenBank) (28)
mRNA (35)
RefSeq (153)

Manage Filters

Top Organisms [Tree]

Homo sapiens	*(18)*
Mus musculus	*(17)*
Callithrix jacchus	*(8)*
Bos taurus	*(6)*
Rattus norvegicus	*(5)*
All other taxa	*(129)*

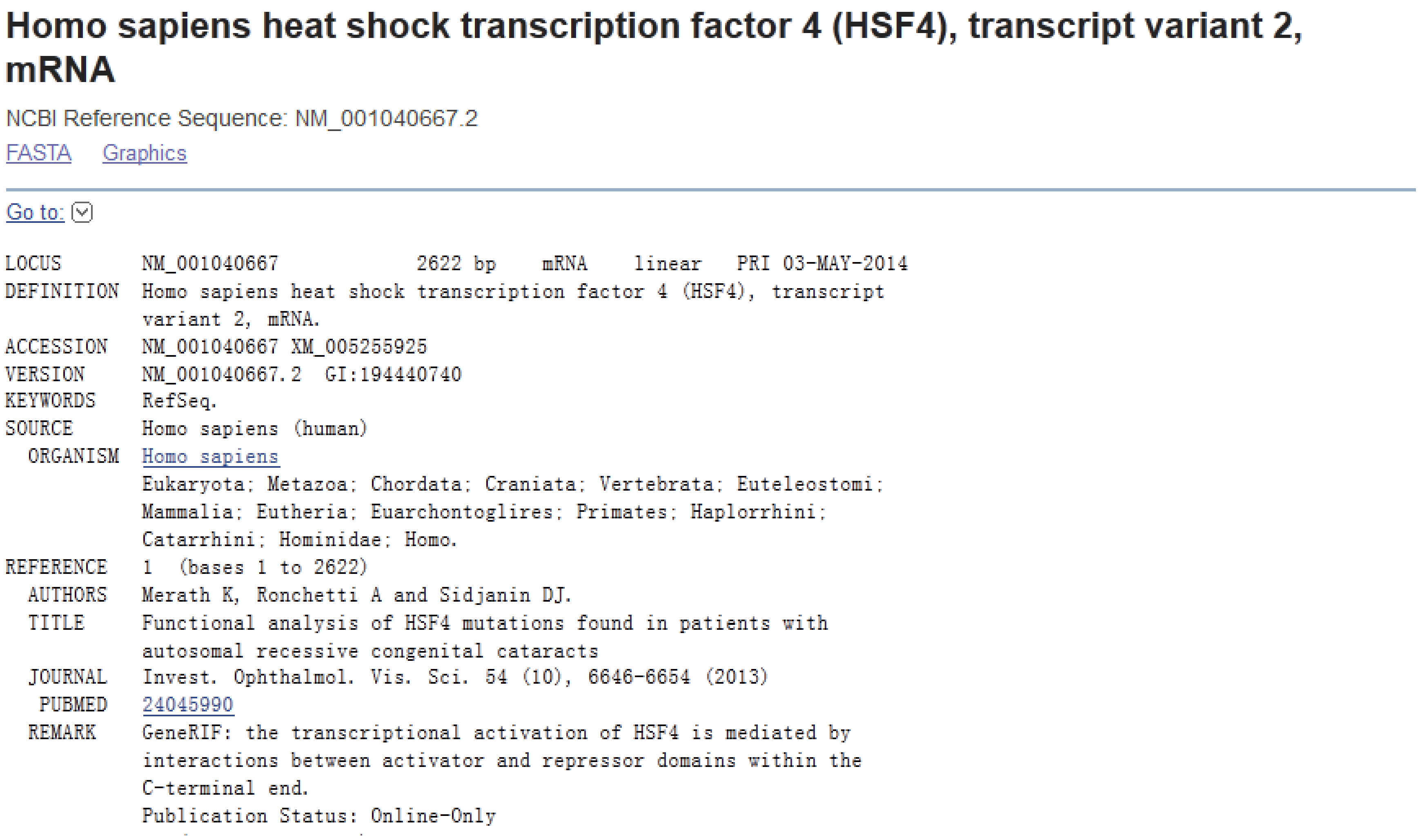
More...

Find related data

Database: Select

Getting sequence from the GenBank database

Data format of nucleic acid (I)



Homo sapiens heat shock transcription factor 4 (HSF4), transcript variant 2, mRNA

NCBI Reference Sequence: NM_001040667.2

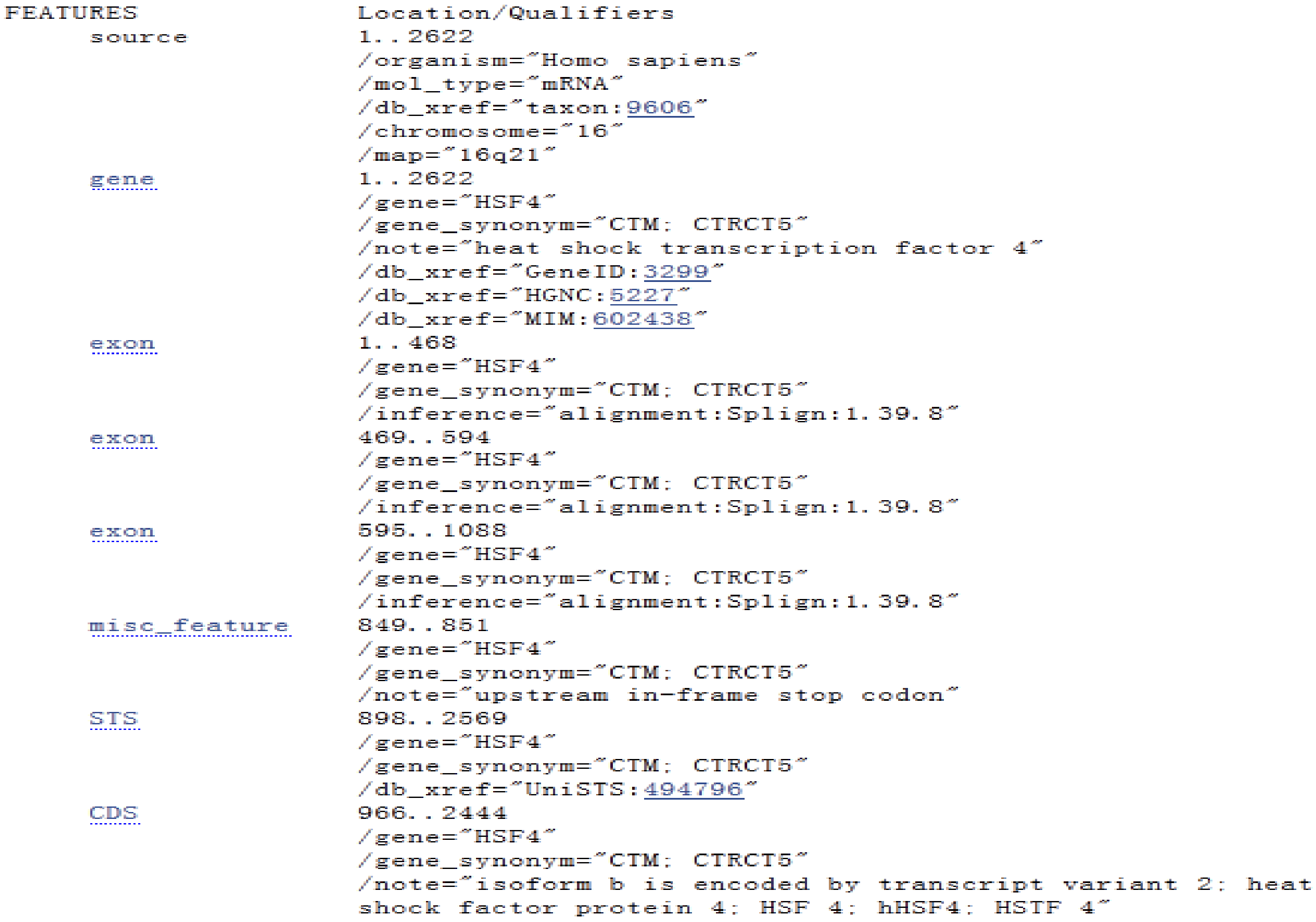
FASTA   Graphics

Go to: ⌄

```
LOCUS       NM_001040667            2622 bp    mRNA    linear   PRI 03-MAY-2014
DEFINITION  Homo sapiens heat shock transcription factor 4 (HSF4), transcript
            variant 2, mRNA.
ACCESSION   NM_001040667 XM_005255925
VERSION     NM_001040667.2  GI:194440740
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 2622)
  AUTHORS   Merath K, Ronchetti A and Sidjanin DJ.
  TITLE     Functional analysis of HSF4 mutations found in patients with
            autosomal recessive congenital cataracts
  JOURNAL   Invest. Ophthalmol. Vis. Sci. 54 (10), 6646-6654 (2013)
   PUBMED   24045990
  REMARK    GeneRIF: the transcriptional activation of HSF4 is mediated by
            interactions between activator and repressor domains within the
            C-terminal end.
            Publication Status: Online-Only
```

Getting sequence from the GenBank database

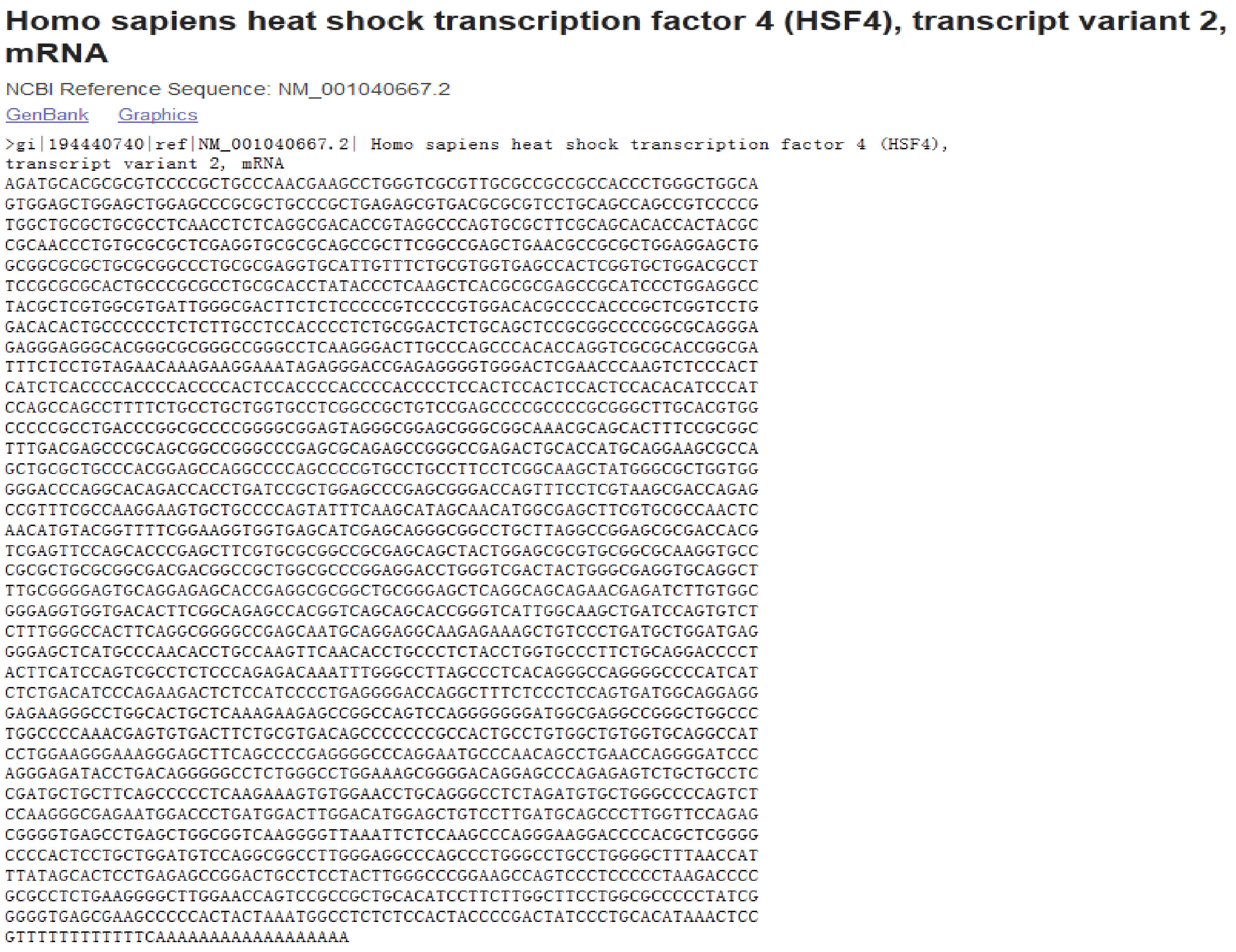
Data format of nucleic acid (II)



```
FEATURES             Location/Qualifiers
     source          1..2622
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="16"
                     /map="16q21"
     gene            1..2622
                     /gene="HSF4"
                     /gene_synonym="CTM; CTRCT5"
                     /note="heat shock transcription factor 4"
                     /db_xref="GeneID:3299"
                     /db_xref="HGNC:5227"
                     /db_xref="MIM:602438"
     exon            1..468
                     /gene="HSF4"
                     /gene_synonym="CTM; CTRCT5"
                     /inference="alignment:Splign:1.39.8"
     exon            469..594
                     /gene="HSF4"
                     /gene_synonym="CTM; CTRCT5"
                     /inference="alignment:Splign:1.39.8"
     exon            595..1088
                     /gene="HSF4"
                     /gene_synonym="CTM; CTRCT5"
                     /inference="alignment:Splign:1.39.8"
     misc_feature    849..851
                     /gene="HSF4"
                     /gene_synonym="CTM; CTRCT5"
                     /note="upstream in-frame stop codon"
     STS             898..2569
                     /gene="HSF4"
                     /gene_synonym="CTM; CTRCT5"
                     /db_xref="UniSTS:494796"
     CDS             966..2444
                     /gene="HSF4"
                     /gene_synonym="CTM; CTRCT5"
                     /note="isoform b is encoded by transcript variant 2; heat
                     shock factor protein 4; HSF 4; hHSF4; HSTF 4"
```

Getting sequence from the GenBank database

Data format of nucleic acid (III)



Homo sapiens heat shock transcription factor 4 (HSF4), transcript variant 2, mRNA

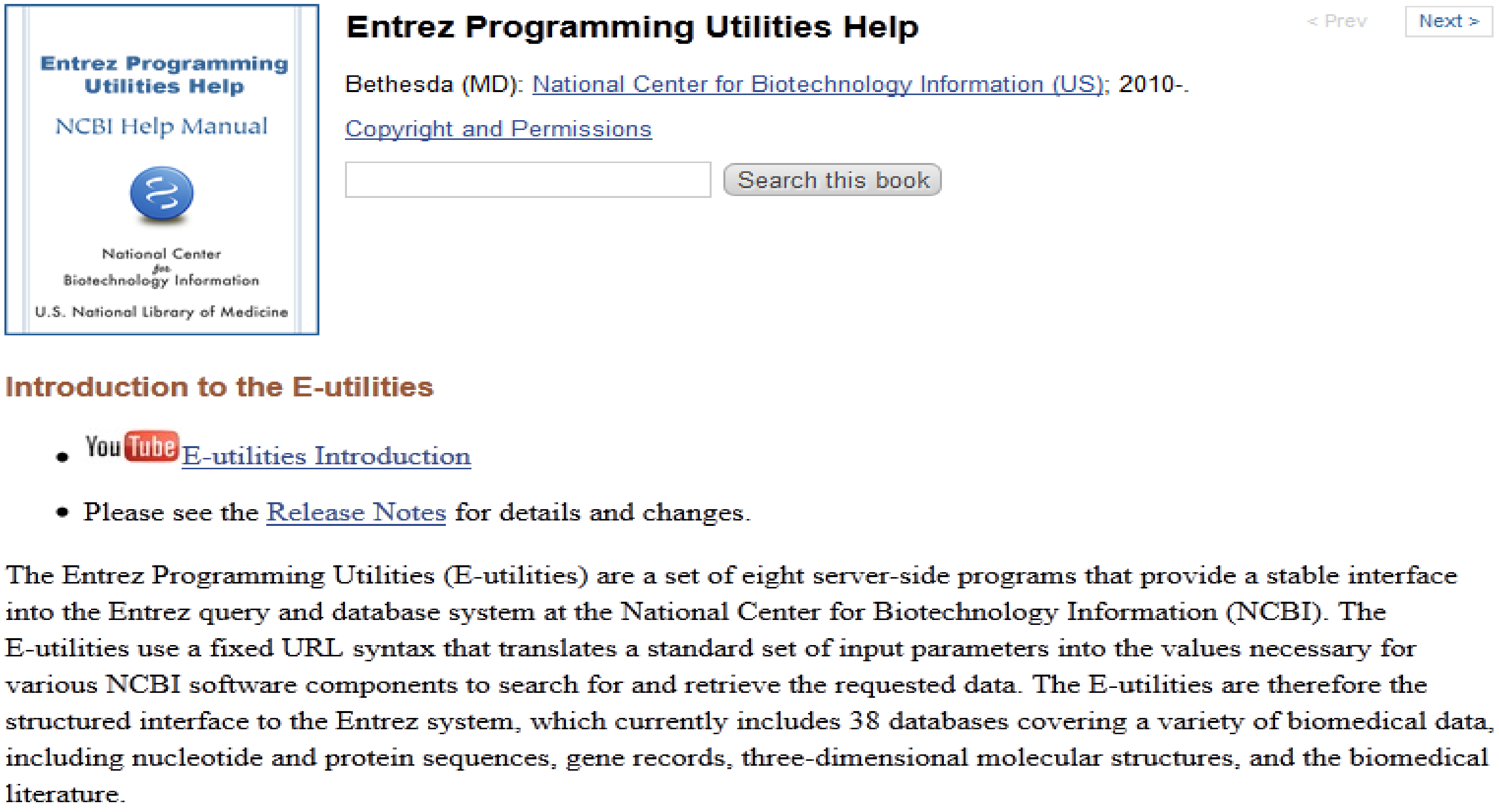
NCBI Reference Sequence: NM_001040667.2

GenBank Graphics

```
>gi|194440740|ref|NM_001040667.2| Homo sapiens heat shock transcription factor 4 (HSF4),
transcript variant 2, mRNA
AGATGCACGCGCGTCCCCGCTGCCCAACGAAGCCTGGGTCGCGTTGCGCCGCCGCCACCCTGGGCTGGCA
GTGGAGCTGGAGCTGGAGCCCGCGCTGCCCGCTGAGAGCGTGACGCGCGTCCTGCAGCCAGCCGTCCCCG
TGGCTGCGCTGCGCCTCAACCTCTCAGGCGACACCGTAGGCCCAGTGCGCTTCGCAGCACACCACTACGC
CGCAACCCTGTGCGCGCTCGAGGTGCGCGCAGCCGCTTCGGCCGAGCTGAACGCCGCGCTGGAGGAGCTG
GCGGCGCGCTGCGCGGCCCTGCGCGAGGTGCATTGTTTCTGCGTGGTGAGCCACTCGGTGCTGGACGCCT
TCCGCGCGCACTGCCCGCGCCTGCGCACCTATACCCTCAAGCTCACGCGCGAGCCGCATCCCTGGAGGCC
TACGCTCGTGGCGTGATTGGGCGACTTCTCTCCCCCGTCCCCGTGGACACGCCCCACCCGCTCGGTCCTG
GACACACTGCCCCCCTCTCTTGCCTCCACCCCTCTGCGGACTCTGCAGCTCCGCGGCCCCGGCGCAGGGA
GAGGGAGGGCACGGGCGCGGGCCGGGCCTCAAGGGACTTGCCCAGCCCACACCAGGTCGCGCACCGGCGA
TTTCTCCTGTAGAACAAAGAAGGAAATAGAGGGACCGAGAGGGGTGGGACTCGAACCCAAGTCTCCCACT
CATCTCACCCCACCCCACCCCACTCCACCCCACCCCACCCCTCCACTCCACTCCACTCCACACATCCCAT
CCAGCCAGCCTTTTCTGCCTGCTGGTGCCTCGGCCGCTGTCCGAGCCCCGCCCCGCGGGCTTGCACGTGG
CCCCCGCCTGACCCGGCGCCCCGGGGCGGAGTAGGGCGGAGCGGGCGGCAAACGCAGCACTTTCCGCGGC
TTTGACGAGCCCGCAGCGGCCGGGCCCGAGCGCAGAGCCGGGCCGAGACTGCACCATGCAGGAAGCGCCA
GCTGCGCTGCCCACGGAGCCAGGCCCCAGCCCCGTGCCTGCCTTCCTCGGCAAGCTATGGGCGCTGGTGG
GGGACCCAGGCACAGACCACCTGATCCGCTGGAGCCCGAGCGGGACCAGTTTCCTCGTAAGCGACCAGAG
CCGTTTCGCCAAGGAAGTGCTGCCCCAGTATTTCAAGCATAGCAACATGGCGAGCTTCGTGCGCCAACTC
AACATGTACGGTTTTCGGAAGGTGGTGAGCATCGAGCAGGGCGGCCTGCTTAGGCCGGAGCGCGACCACG
TCGAGTTCCAGCACCCGAGCTTCGTGCGCGGCCGCGAGCAGCTACTGGAGCGCGTGCGGCGCAAGGTGCC
CGCGCTGCGCGGCGACGACGGCCGCTGGCGCCCGGAGGACCTGGGTCGACTACTGGGCGAGGTGCAGGCT
TTGCGGGGAGTGCAGGAGAGCACCGAGGCGCGGCTGCGGGAGCTCAGGCAGCAGAACGAGATCTTGTGGC
GGGAGGTGGTGACACTTCGGCAGAGCCACGGTCAGCAGCACCGGGTCATTGGCAAGCTGATCCAGTGTCT
CTTTGGGCCACTTCAGGCGGGGCCGAGCAATGCAGGAGGCAAGAGAAAGCTGTCCCTGATGCTGGATGAG
GGGAGCTCATGCCCAACACCTGCCAAGTTCAACACCTGCCCTCTACCTGGTGCCCTTCTGCAGGACCCCT
ACTTCATCCAGTCGCCTCTCCCAGAGACAAATTTGGGCCTTAGCCCTCACAGGGCCAGGGGCCCCATCAT
CTCTGACATCCCAGAAGACTCTCCATCCCCTGAGGGGACCAGGCTTTCTCCCTCCAGTGATGGCAGGAGG
GAGAAGGGCCTGGCACTGCTCAAAGAAGAGCCGGCCAGTCCAGGGGGGGATGGCGAGGCCGGGCTGGCCC
TGGCCCCAAACGAGTGTGACTTCTGCGTGACAGCCCCCCCGCCACTGCCTGTGGCTGTGGTGCAGGCCAT
CCTGGAAGGGAAAGGGAGCTTCAGCCCCGAGGGGCCCAGGAATGCCCAACAGCCTGAACCAGGGGATCCC
AGGGAGATACCTGACAGGGGGCCTCTGGGCCTGGAAAGCGGGGACAGGAGCCCAGAGAGTCTGCTGCCTC
CGATGCTGCTTCAGCCCCCTCAAGAAAGTGTGGAACCTGCAGGGCCTCTAGATGTGCTGGGCCCCAGTCT
CCAAGGGCGAGAATGGACCCTGATGGACTTGGACATGGAGCTGTCCTTGATGCAGCCCTTGGTTCCAGAG
CGGGGTGAGCCTGAGCTGGCGGTCAAGGGGTTAAATTCTCCAAGCCCAGGGAAGGACCCCACGCTCGGGG
CCCCACTCCTGCTGGATGTCCAGGCGGCCTTGGGAGGCCCAGCCCTGGGCCTGCCTGGGGCTTTAACCAT
TTATAGCACTCCTGAGAGCCGGACTGCCTCCTACTTGGGCCCGGAAGCCAGTCCCTCCCCCTAAGACCCC
GCGCCTCTGAAGGGGCTTGGAACCAGTCCGCCGCTGCACATCCTTCTTGGCTTCCTGGCGCCCCCTATCG
GGGGTGAGCGAAGCCCCCACTACTAAATGGCCTCTCTCCACTACCCCGACTATCCCTGCACATAAACTCC
GTTTTTTTTTTTTCAAAAAAAAAAAAAAAAAA
```

Getting sequence from the GenBank database

http://www.ncbi.nlm.nih.gov/books/NBK25501/


Entrez Programming Utilities Help
NCBI Help Manual
National Center for Biotechnology Information
U.S. National Library of Medicine
< Prev
Next >
Search this book
You Tube

Entrez Programming Utilities Help

Bethesda (MD): National Center for Biotechnology Information (US); 2010-.

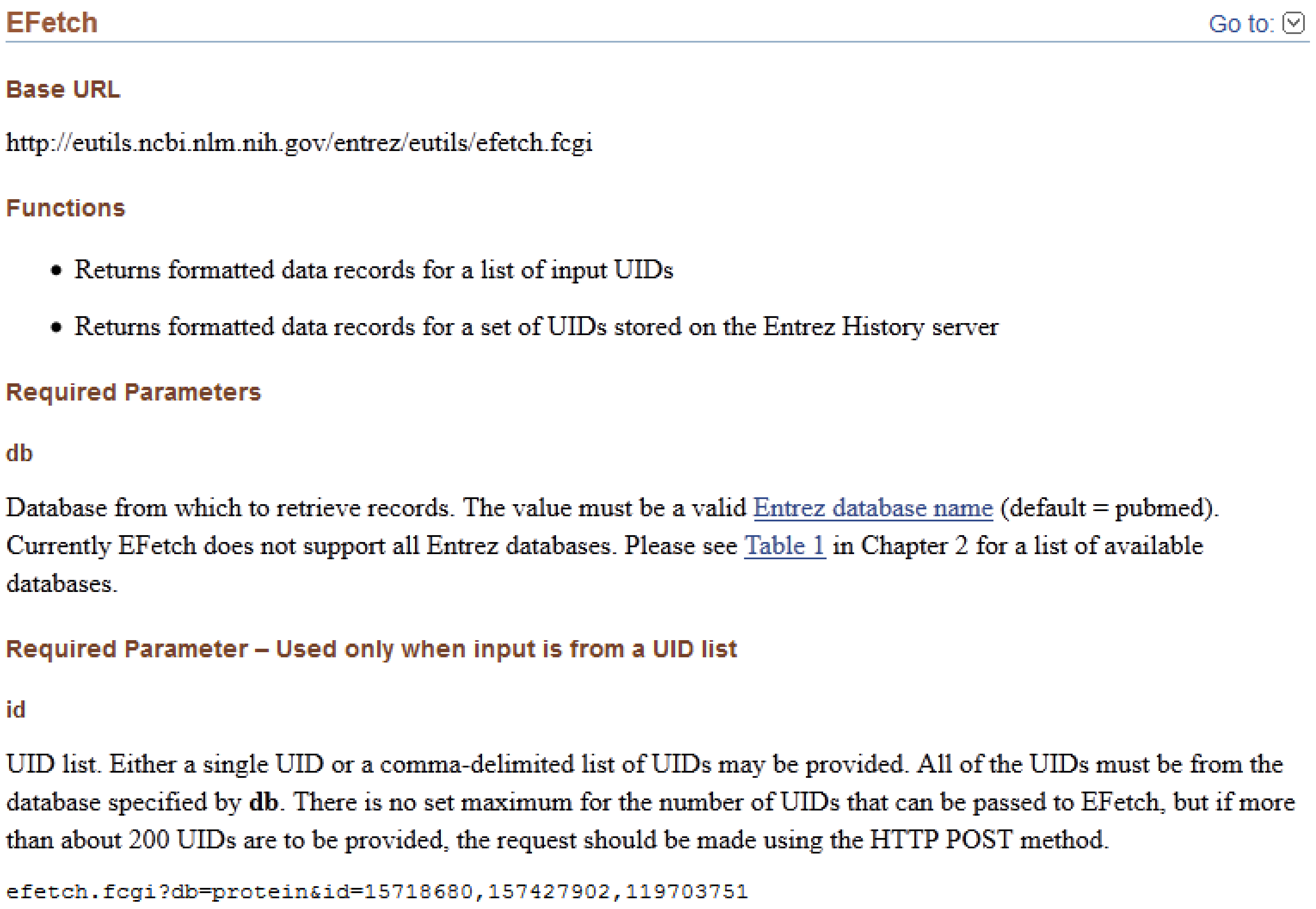
Copyright and Permissions

Introduction to the E-utilities

- E-utilities Introduction
- Please see the Release Notes for details and changes.

The Entrez Programming Utilities (E-utilities) are a set of eight server-side programs that provide a stable interface into the Entrez query and database system at the National Center for Biotechnology Information (NCBI). The E-utilities use a fixed URL syntax that translates a standard set of input parameters into the values necessary for various NCBI software components to search for and retrieve the requested data. The E-utilities are therefore the structured interface to the Entrez system, which currently includes 38 databases covering a variety of biomedical data, including nucleotide and protein sequences, gene records, three-dimensional molecular structures, and the biomedical literature.

Getting sequence from the GenBank database



EFetch

Go to: ⌄

Base URL

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi

Functions

- Returns formatted data records for a list of input UIDs
- Returns formatted data records for a set of UIDs stored on the Entrez History server

Required Parameters

db

Database from which to retrieve records. The value must be a valid Entrez database name (default = pubmed). Currently EFetch does not support all Entrez databases. Please see Table 1 in Chapter 2 for a list of available databases.

Required Parameter – Used only when input is from a UID list

id

UID list. Either a single UID or a comma-delimited list of UIDs may be provided. All of the UIDs must be from the database specified by **db**. There is no set maximum for the number of UIDs that can be passed to EFetch, but if more than about 200 UIDs are to be provided, the request should be made using the HTTP POST method.

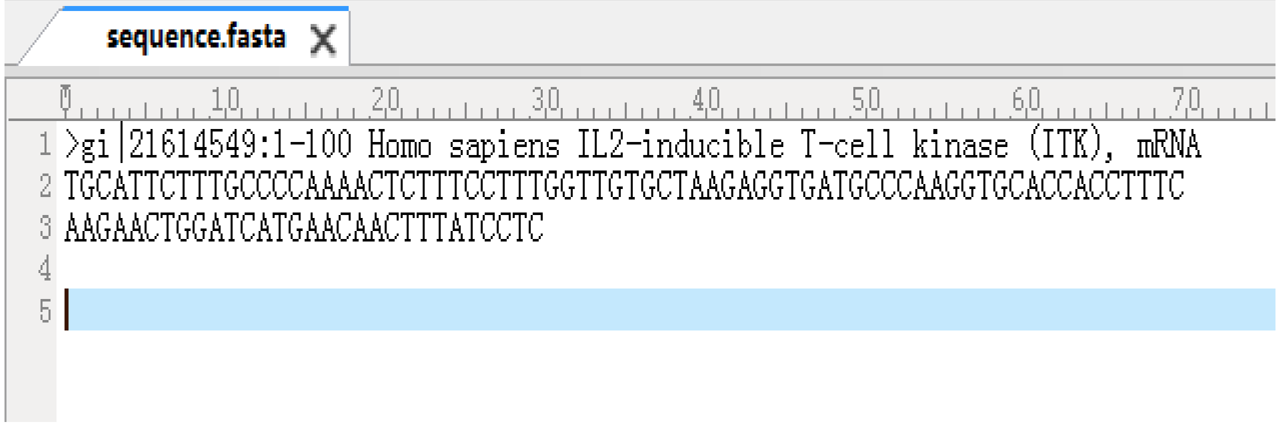
```
efetch.fcgi?db=protein&id=15718680,157427902,119703751
```

Getting sequence from the GenBank database

Nucleotide/Nuccore

Fetch the first 100 bases of the plus strand of GI 21614549 in FASTA format:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nuccore&id=21614549&strand=1&seq_start=1&seq_stop=100&rettype=fasta&retmode=text



```
>gi|21614549:1-100 Homo sapiens IL2-inducible T-cell kinase (ITK), mRNA
TGCATTCTTTGCCCCAAAACTCTTTCCTTTGGTTGTGCTAAGAGGTGATGCCCAAGGTGCACCACCTTTC
AAGAACTGGATCATGAACAACTTTATCCTC
```

PART II Analysis of nucleic acid sequence

Analysis of nucleic acid sequence

(1) Getting basic information of a nucleic acid sequence

(2) Primer design

(3) Tow sequences alignment

(4) Multi sequences alignment

(5) Finding open reading frame of a nucleic acid sequence

(6) Gene prediction

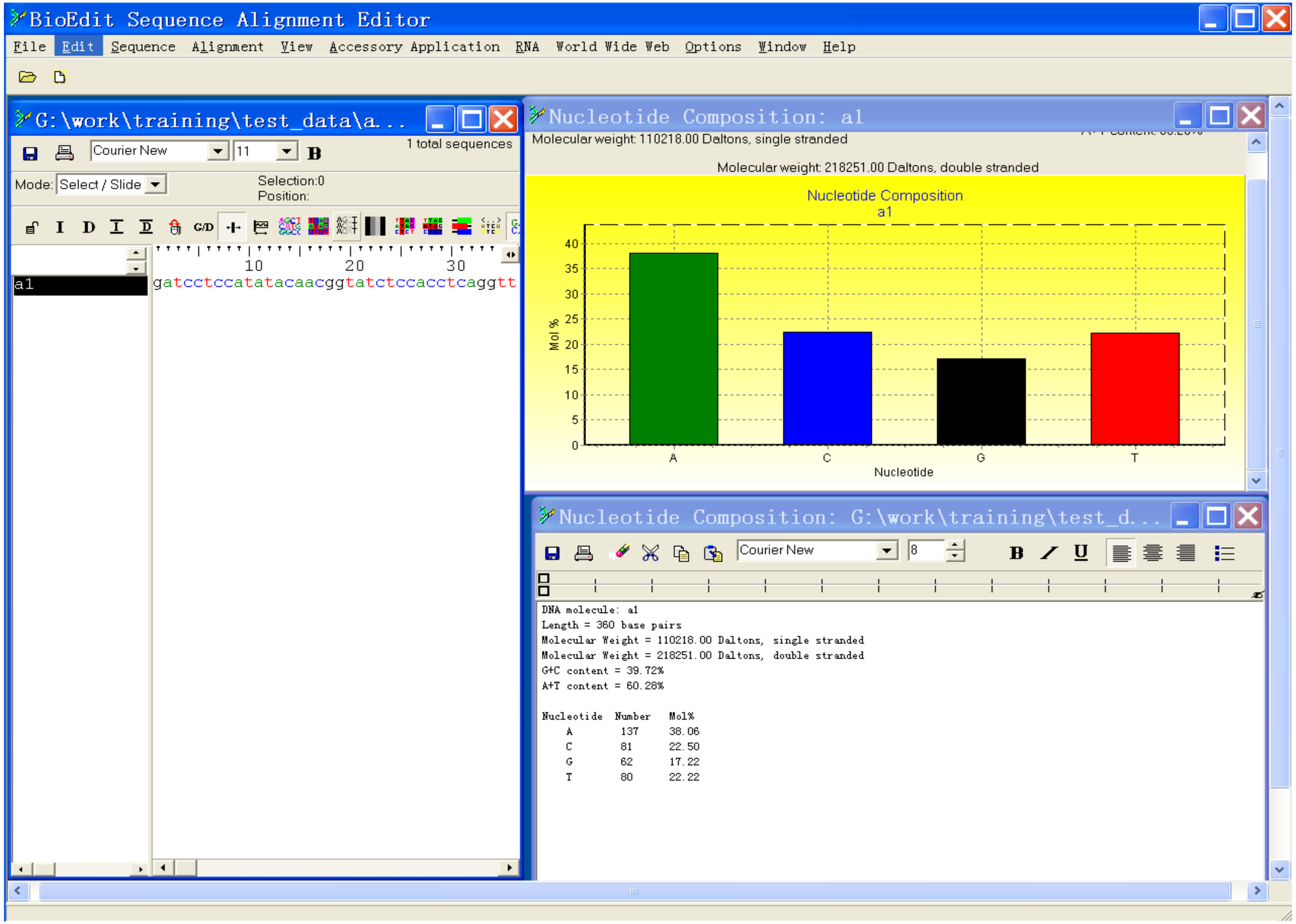
(7) Sequence localization in genome

(8) Sequence assembly

Obtain basic information of nucleic acid sequence

Software: BioEdit (http://www.mbio.ncsu.edu/bioedit/bioedit.html)

menu: File -> open -> sequence -> nucleic acid -> nucleotide composition


BioEdit Sequence Alignment Editor
File Edit Sequence Alignment View Accessory Application RNA World Wide Web Options Window Help
G:\work\training\test_data\a...
Courier New
11
1 total sequences
Mode: Select / Slide
Selection:0
Position:
10
20
30
a1
gatcctccatatacaacggtatctccacctcaggtt
Nucleotide Composition: a1
Molecular weight: 110218.00 Daltons, single stranded
Molecular weight: 218251.00 Daltons, double stranded
Nucleotide Composition
a1
Mol %
40
35
30
25
20
15
10
5
0
A
C
G
T
Nucleotide
Nucleotide Composition: G:\work\training\test_d...
Courier New
8

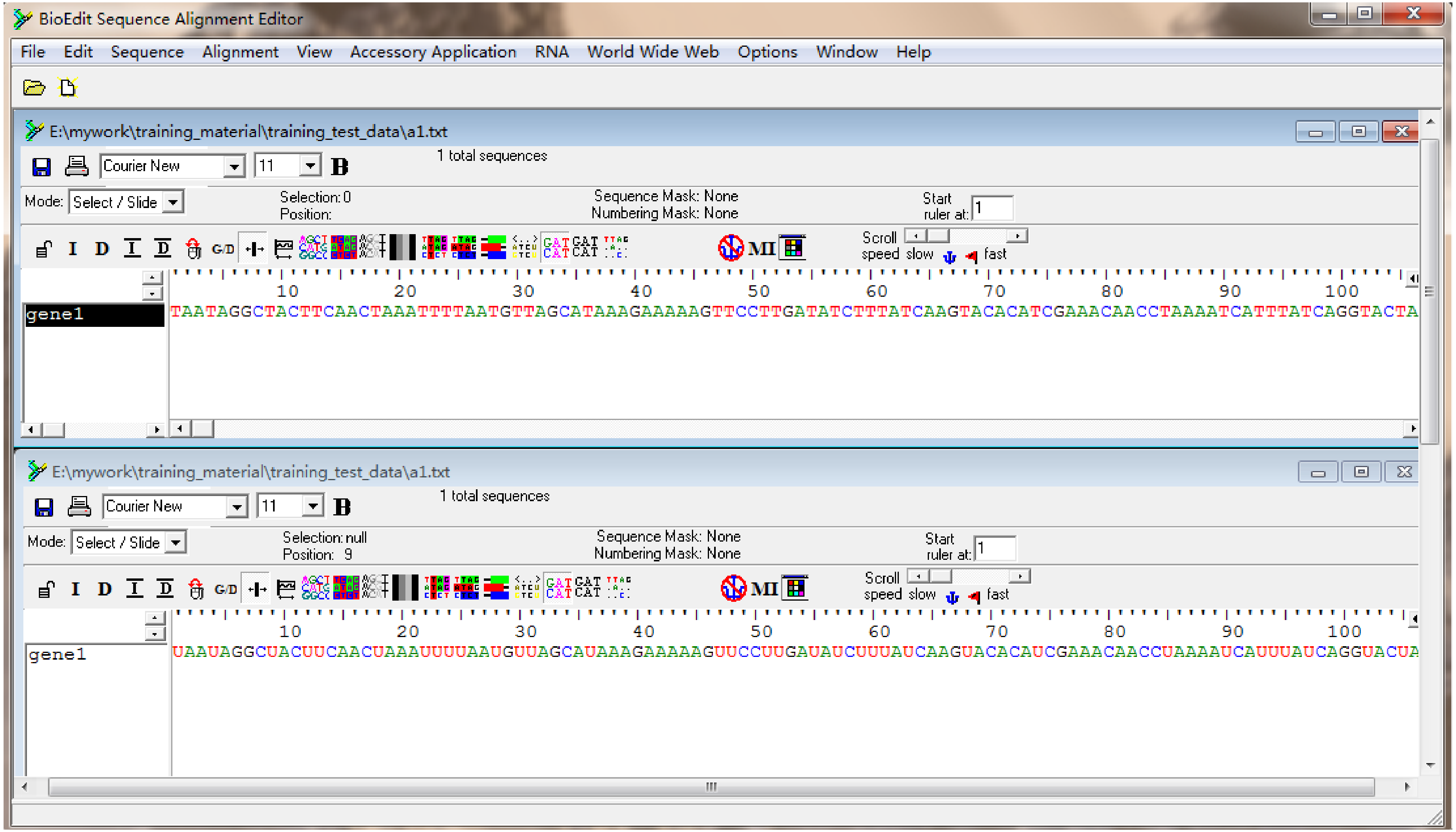
```
DNA molecule: a1
Length = 360 base pairs
Molecular Weight = 110218.00 Daltons, single stranded
Molecular Weight = 218251.00 Daltons, double stranded
G+C content = 39.72%
A+T content = 60.28%

Nucleotide  Number   Mol%
    A         137    38.06
    C         81     22.50
    G         62     17.22
    T         80     22.22
```

Basic transition of nucleic acid sequence

- DNA -> RNA
- Sequence -> Reverse complement
- DNA -> protein

Basic transition of nucleic acid sequence

menu: File -> open -> sequence -> nucleic acid -> DNA-RNA



E:\mywork\training_material\training_test_data\a1.txt

Courier New 11

1 total sequences

Mode: Select / Slide

Selection: 0
Position:

Sequence Mask: None
Numbering Mask: None

Start ruler at: 1

```
gene1 TAATAGGCTACTTCAACTAAATTTTAATGTTAGCATAAAGAAAAAGTTCCTTGATATCTTTATCAAGTACACATCGAAACAACCTAAAATCATTTATCAGGTACTA
```

E:\mywork\training_material\training_test_data\a1.txt

Courier New 11

1 total sequences

Mode: Select / Slide

Selection: null
Position: 9

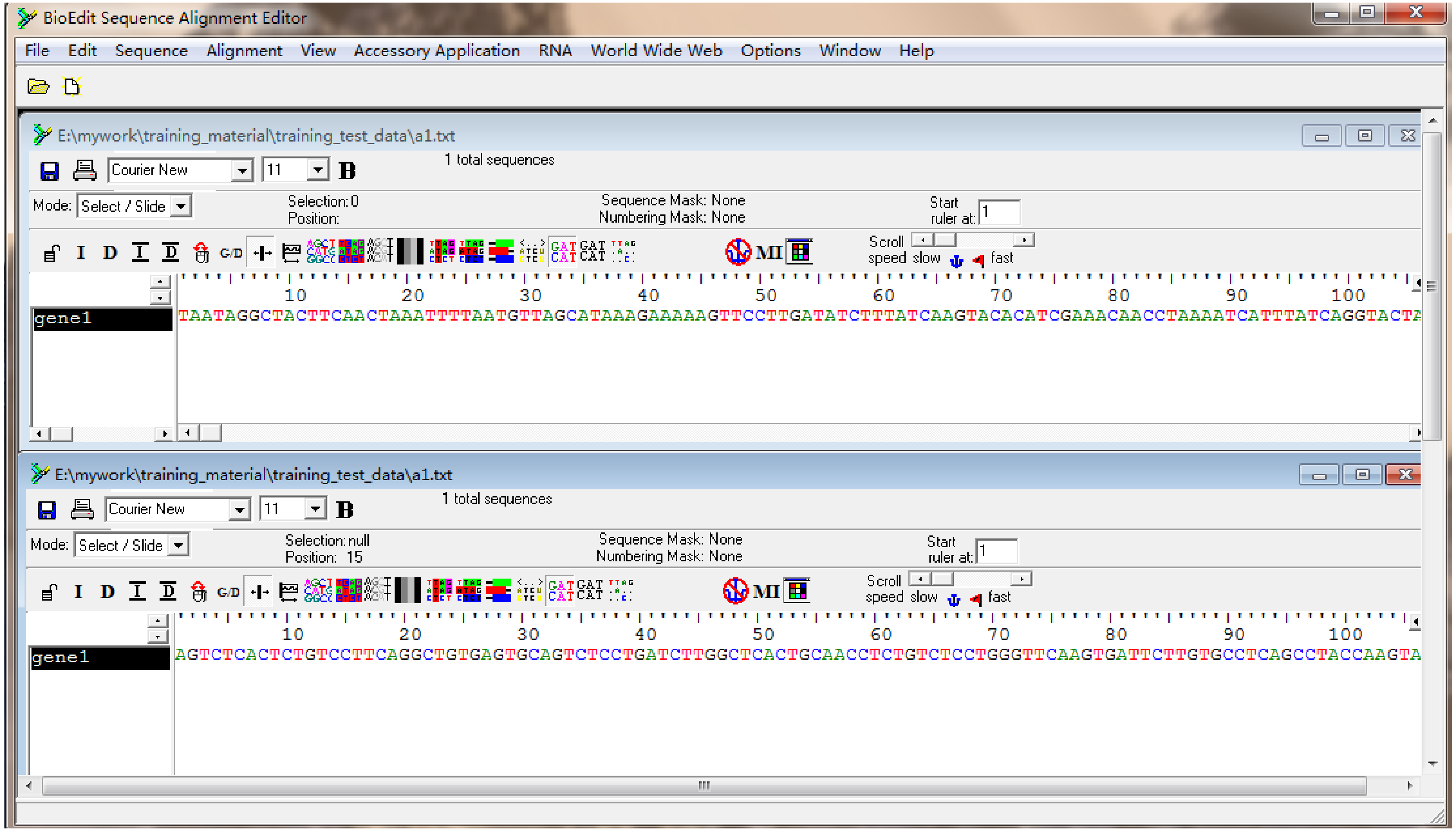
Sequence Mask: None
Numbering Mask: None

Start ruler at: 1

```
gene1 UAAUAGGCUACUUCAACUAAAUUUUAAUGUUAGCAUAAAGAAAAAGUUCCUUGAUAUCUUUAUCAAGUACACAUCGAAACAACCUAAAAUCAUUUAUCAGGUACUA
```

Basic transition of nucleic acid sequence

menu: File -> open -> sequence -> nucleic acid -> Reverse Complement



E:\mywork\training_material\training_test_data\a1.txt

Courier New 11 B — 1 total sequences

Mode: Select / Slide — Selection: 0 — Position: — Sequence Mask: None — Numbering Mask: None — Start ruler at: 1

```
gene1  TAATAGGCTACTTCAACTAAATTTTAATGTTAGCATAAAGAAAAAGTTCCTTGATATCTTTATCAAGTACACATCGAAACAACCTAAAATCATTTATCAGGTACTA
```

E:\mywork\training_material\training_test_data\a1.txt

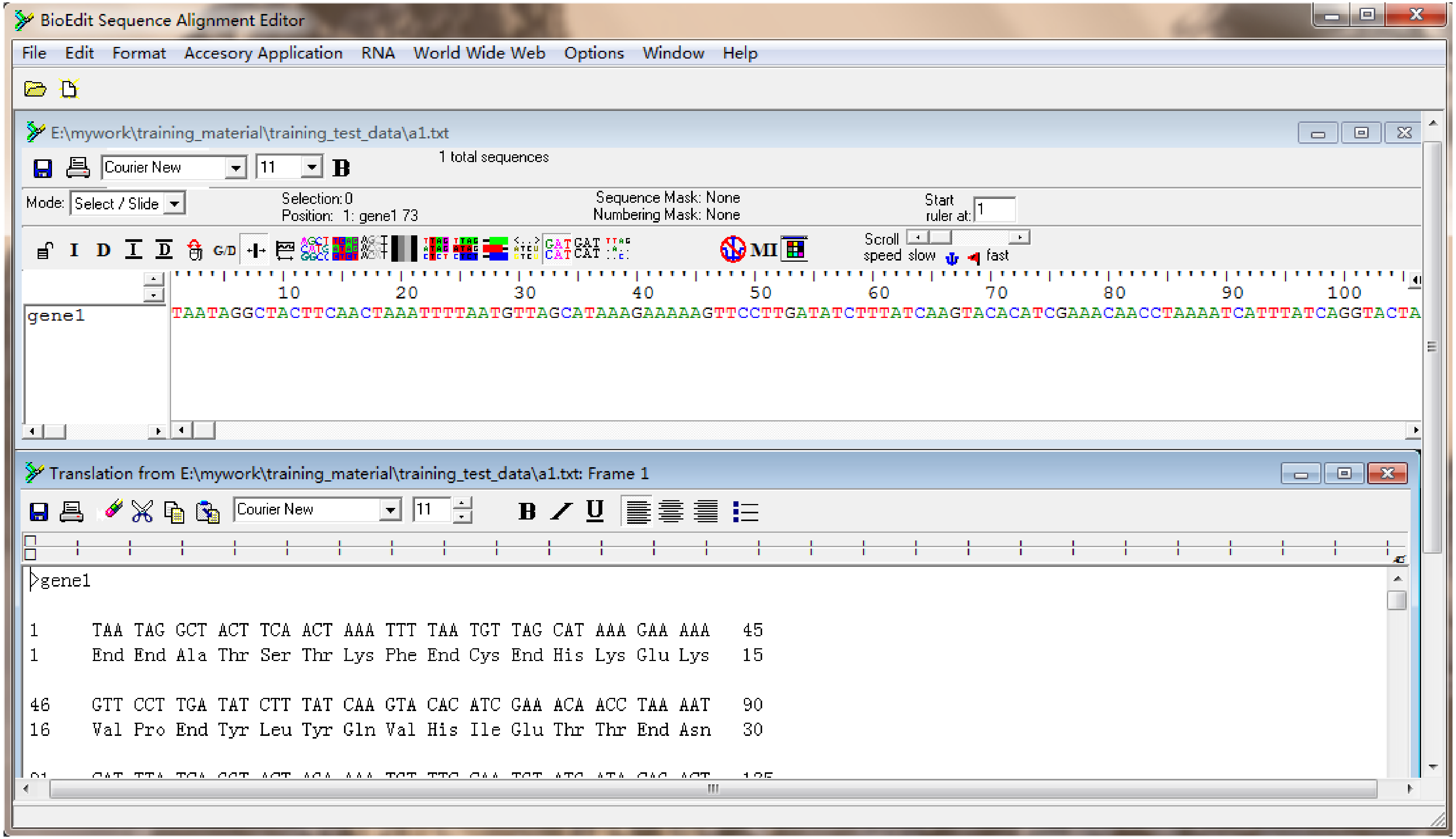
Courier New 11 B — 1 total sequences

Mode: Select / Slide — Selection: null — Position: 15 — Sequence Mask: None — Numbering Mask: None — Start ruler at: 1

```
gene1  AGTCTCACTCTGTCCTTCAGGCTGTGAGTGCAGTCTCCTGATCTTGGCTCACTGCAACCTCTGTCTCCTGGGTTCAAGTGATTCTTGTGCCTCAGCCTACCAAGTA
```

Basic transition of nucleic acid sequence

menu: File -> open -> sequence -> nucleic acid -> translate



E:\mywork\training_material\training_test_data\a1.txt

Courier New 11

1 total sequences

Mode: Select / Slide

Selection: 0
Position: 1: gene1 73

Sequence Mask: None
Numbering Mask: None

Start ruler at: 1

```
          10        20        30        40        50        60        70        80        90        100
gene1  TAATAGGCTACTTCAACTAAATTTTAATGTTAGCATAAAGAAAAAGTTCCTTGATATCTTTATCAAGTACACATCGAAACAACCTAAAATCATTTATCAGGTACTA
```

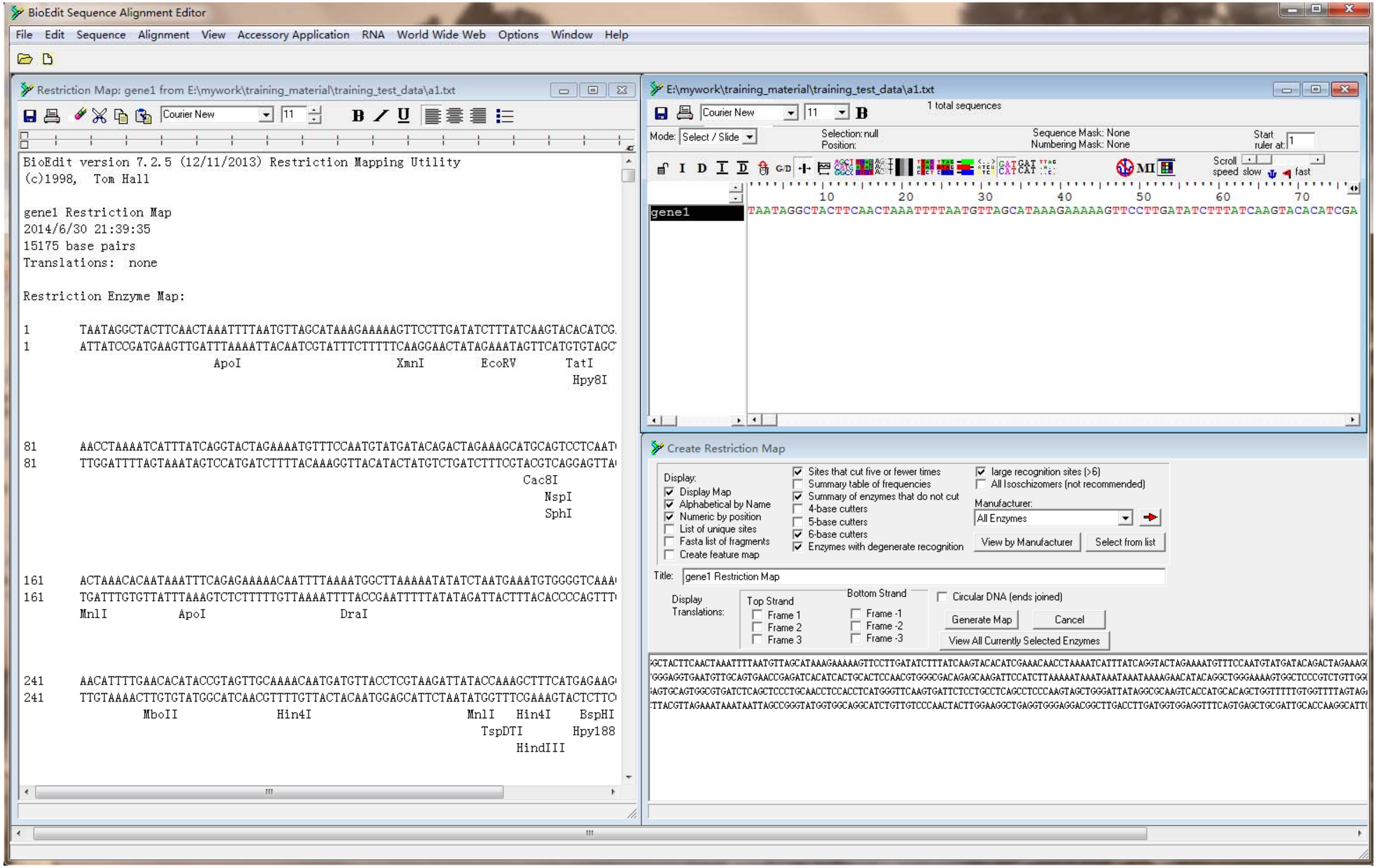
Translation from E:\mywork\training_material\training_test_data\a1.txt: Frame 1

Courier New 11

```
>gene1

1      TAA TAG GCT ACT TCA ACT AAA TTT TAA TGT TAG CAT AAA GAA AAA   45
1      End End Ala Thr Ser Thr Lys Phe End Cys End His Lys Glu Lys   15

46     GTT CCT TGA TAT CTT TAT CAA GTA CAC ATC GAA ACA ACC TAA AAT   90
16     Val Pro End Tyr Leu Tyr Gln Val His Ile Glu Thr Thr End Asn   30

91     CAT TTA TCA GGT ACT ACA AAA TGT TTG GAA TGT ATG ATA CAG ACT   135
```

Analysis of enzyme mapping

menu: File -> open -> sequence -> nucleic acid -> Restriction Map


BioEdit Sequence Alignment Editor
File Edit Sequence Alignment View Accessory Application RNA World Wide Web Options Window Help
Restriction Map: gene1 from E:\mywork\training_material\training_test_data\a1.txt
Courier New
11
E:\mywork\training_material\training_test_data\a1.txt
Courier New
11
Create Restriction Map

```
BioEdit version 7.2.5 (12/11/2013) Restriction Mapping Utility
(c)1998,  Tom Hall

gene1 Restriction Map
2014/6/30 21:39:35
15175 base pairs
Translations:  none

Restriction Enzyme Map:

1       TAATAGGCTACTTCAACTAAATTTTAATGTTAGCATAAAGAAAAAGTTCCTTGATATCTTTATCAAGTACACATCG
1       ATTATCCGATGAAGTTGATTTAAAATTACAATCGTATTTCTTTTTCAAGGAACTATAGAAATAGTTCATGTGTAGC
                            ApoI                       XmnI         EcoRV       TatI
                                                                                 Hpy8I


81      AACCTAAAATCATTTATCAGGTACTAGAAAATGTTTCCAATGTATGATACAGACTAGAAAGCATGCAGTCCTCAAT
81      TTGGATTTTAGTAAATAGTCCATGATCTTTTACAAAGGTTACATACTATGTCTGATCTTTCGTACGTCAGGAGTTA
                                                                          Cac8I
                                                                             NspI
                                                                             SphI


161     ACTAAACACAATAAATTTCAGAGAAAAACAATTTTAAAATGGCTTAAAAATATATCTAATGAAATGTGGGGTCAAA
161     TGATTTGTGTTATTTAAAGTCTCTTTTTGTTAAAATTTTACCGAATTTTTATATAGATTACTTTACACCCCAGTTT
        MnlI           ApoI                    DraI


241     AACATTTTGAACACATACCGTAGTTGCAAAACAATGATGTTACCTCGTAAGATTATACCAAAGCTTTCATGAGAAG
241     TTGTAAAACTTGTGTATGGCATCAACGTTTTGTTACTACAATGGAGCATTCTAATATGGTTTCGAAAGTACTCTTC
                  MboII              Hin4I                       MnlI   Hin4I    BspHI
                                                                    TspDTI      Hpy188
                                                                         HindIII
```

1 total sequences

Mode: Select / Slide

Selection: null
Position:

Sequence Mask: None
Numbering Mask: None

Start ruler at: 1

Scroll speed slow fast

```
             10        20        30        40        50        60        70
gene1   TAATAGGCTACTTCAACTAAATTTTAATGTTAGCATAAAGAAAAAGTTCCTTGATATCTTTATCAAGTACACATCGA
```

Create Restriction Map

Display:
- ☑ Display Map
- ☑ Alphabetical by Name
- ☑ Numeric by position
- □ List of unique sites
- □ Fasta list of fragments
- □ Create feature map

- ☑ Sites that cut five or fewer times
- □ Summary table of frequencies
- ☑ Summary of enzymes that do not cut
- □ 4-base cutters
- □ 5-base cutters
- ☑ 6-base cutters
- ☑ Enzymes with degenerate recognition

- ☑ large recognition sites (>6)
- □ All Isoschizomers (not recommended)

Manufacturer: All Enzymes

View by Manufacturer | Select from list

Title: gene1 Restriction Map

Display Translations:

Top Strand
- □ Frame 1
- □ Frame 2
- □ Frame 3

Bottom Strand
- □ Frame -1
- □ Frame -2
- □ Frame -3

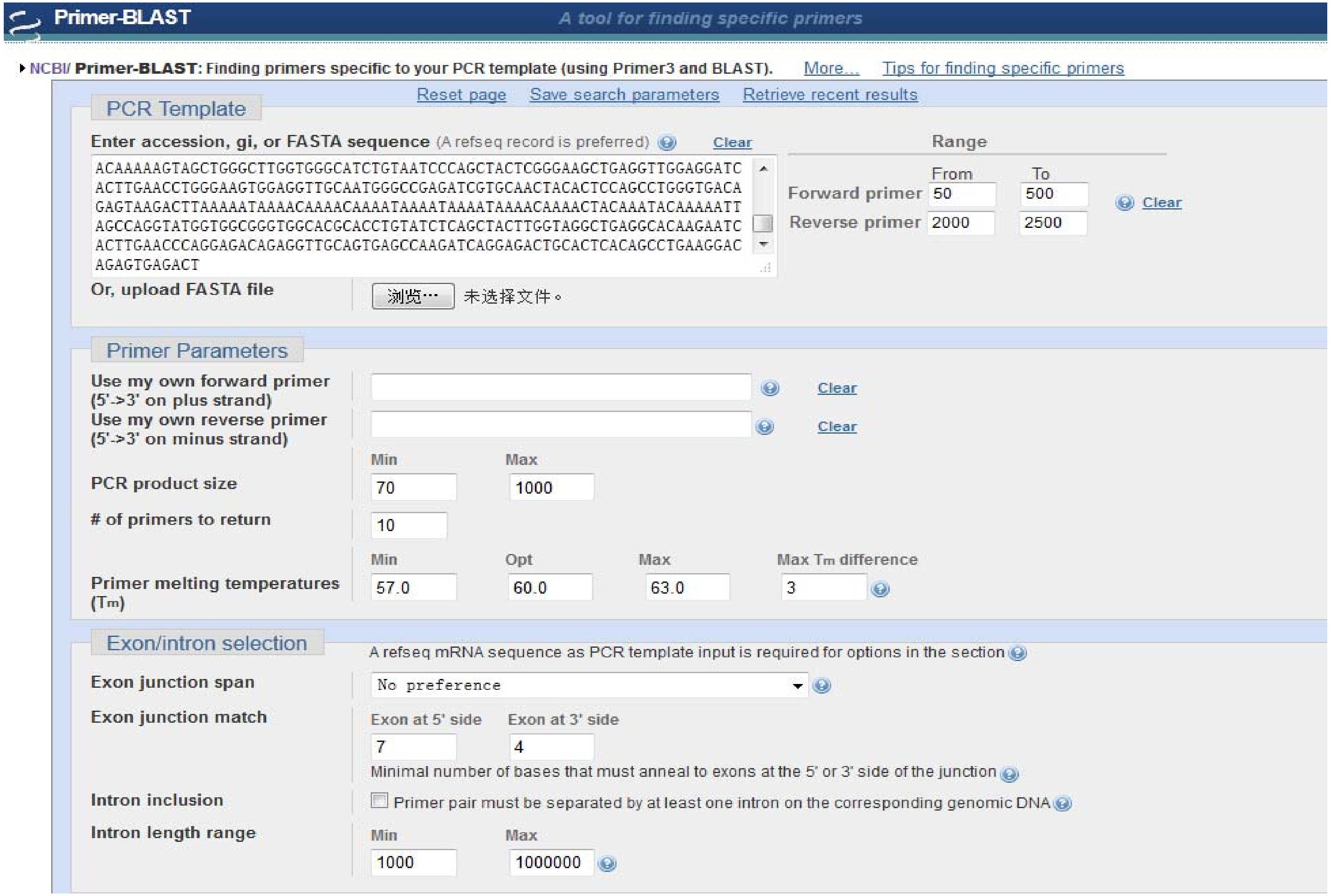
□ Circular DNA (ends joined)

Generate Map | Cancel | View All Currently Selected Enzymes

```
GCTACTTCAACTAAATTTTAATGTTAGCATAAAGAAAAAGTTCCTTGATATCTTTATCAAGTACACATCGAAACAACCTAAAATCATTTATCAGGTACTAGAAAATGTTTCCAATGTATGATACAGACTAGAAAG
GGGAGGTGAATGTTGCAGTGAACCGAGATCACATCACTGCACTCCAACGTGGGCGACAGAGCAAGATTCCATCTTAAAAATAAATAAATAAATAAAAGAACATACAGGCTGGGAAAAGTGGCTCCCGTCTGTTGG
AGTGCAGTGGCGTGATCTCAGCTCCCTGCAACCTCCACCTCATGGGTTCAAGTGATTCTCCTGCCTCAGCCTCCCAAGTAGCTGGGATTATAGGCGCAAGTCACCATGCACAGCTGGTTTTTGTGGTTTTAGTAG
TTACGTTAGAAATAAATAATTAGCCGGGTATGGTGGCAGGCATCTGTTGTCCCAACTACTTGGAAGGCTGAGGTGGGAGGACGGCTTGACCTTGATGGTGGAGGTTTCAGTGAGCTGCGATTGCACCAAGGCATT
```

Primer design

NCBI website (http://www.ncbi.nlm.nih.gov/tools/primer-blast/)



Primer-BLAST   *A tool for finding specific primers*

NCBI/ **Primer-BLAST:** Finding primers specific to your PCR template (using Primer3 and BLAST).   More...   Tips for finding specific primers

Reset page   Save search parameters   Retrieve recent results

PCR Template

Enter accession, gi, or FASTA sequence (A refseq record is preferred)   Clear

```
ACAAAAAGTAGCTGGGCTTGGTGGGCATCTGTAATCCCAGCTACTCGGGAAGCTGAGGTTGGAGGATC
ACTTGAACCTGGGAAGTGGAGGTTGCAATGGGCCGAGATCGTGCAACTACACTCCAGCCTGGGTGACA
GAGTAAGACTTAAAAATAAAACAAAACAAAATAAAATAAAATAAAACAAAACTACAAATACAAAAATT
AGCCAGGTATGGTGGCGGGTGGCACGCACCTGTATCTCAGCTACTTGGTAGGCTGAGGCACAAGAATC
ACTTGAACCCAGGAGACAGAGGTTGCAGTGAGCCAAGATCAGGAGACTGCACTCACAGCCTGAAGGAC
AGAGTGAGACT
```

Or, upload FASTA file   浏览…   未选择文件。

Range	From	To
Forward primer	50	500
Reverse primer	2000	2500

Clear

Primer Parameters

Use my own forward primer (5'->3' on plus strand)   Clear

Use my own reverse primer (5'->3' on minus strand)   Clear

	Min	Opt	Max	Max Tm difference
PCR product size	70		1000	
# of primers to return	10			
Primer melting temperatures (Tm)	57.0	60.0	63.0	3

Exon/intron selection

A refseq mRNA sequence as PCR template input is required for options in the section

Exon junction span   No preference

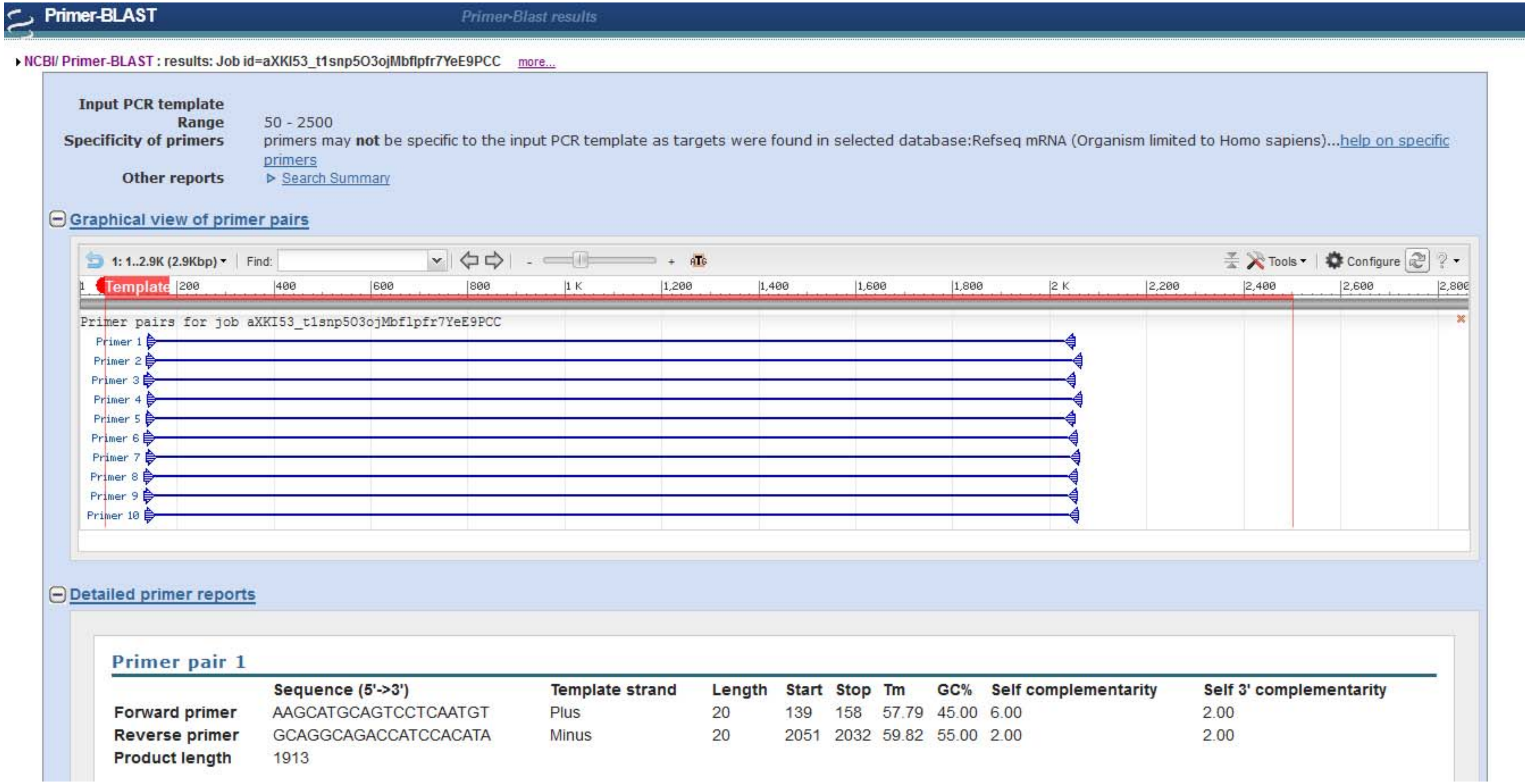
Exon junction match   Exon at 5' side: 7   Exon at 3' side: 4

Minimal number of bases that must anneal to exons at the 5' or 3' side of the junction

Intron inclusion   □ Primer pair must be separated by at least one intron on the corresponding genomic DNA

Intron length range   Min: 1000   Max: 1000000

Primer design



NCBI/ Primer-BLAST : results: Job id=aXKI53_t1snp5O3ojMbflpfr7YeE9PCC more...

Input PCR template	
Range	50 - 2500
Specificity of primers	primers may **not** be specific to the input PCR template as targets were found in selected database:Refseq mRNA (Organism limited to Homo sapiens)...help on specific primers
Other reports	▷ Search Summary

⊖ Graphical view of primer pairs

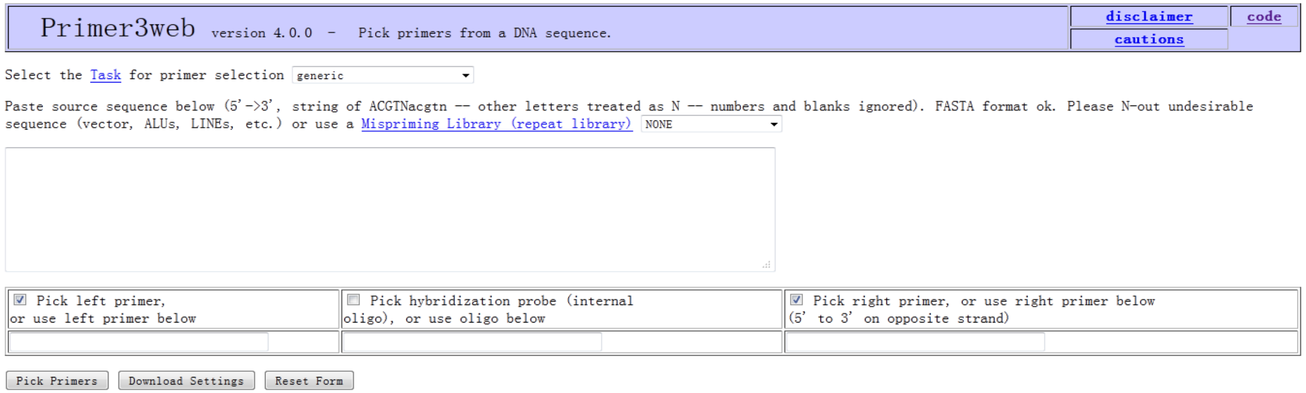
⊖ Detailed primer reports

Primer pair 1

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	AAGCATGCAGTCCTCAATGT	Plus	20	139	158	57.79	45.00	6.00	2.00
Reverse primer	GCAGGCAGACCATCCACATA	Minus	20	2051	2032	59.82	55.00	2.00	2.00
Product length	1913								

Primer design

Primer3 (http://primer3.wi.mit.edu/)



Primer3web version 4.0.0 - Pick primers from a DNA sequence.

Select the Task for primer selection generic

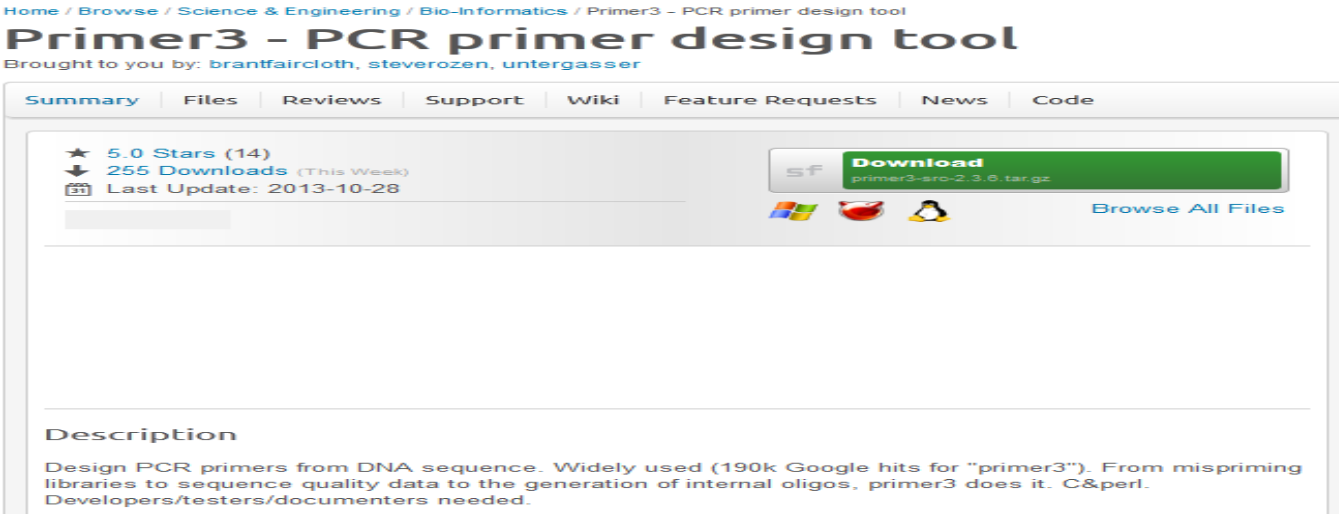
Paste source sequence below (5'->3', string of ACGTNacgtn -- other letters treated as N -- numbers and blanks ignored). FASTA format ok. Please N-out undesirable sequence (vector, ALUs, LINEs, etc.) or use a Mispriming Library (repeat library) NONE

☑ Pick left primer, or use left primer below

□ Pick hybridization probe (internal oligo), or use oligo below

☑ Pick right primer, or use right primer below (5' to 3' on opposite strand)

Primer3 codes (http://sourceforge.net/projects/primer3/)



Home / Browse / Science & Engineering / Bio-Informatics / Primer3 - PCR primer design tool

Primer3 - PCR primer design tool

Brought to you by: brantfaircloth, steverozen, untergasser

★ 5.0 Stars (14)

255 Downloads (This Week)

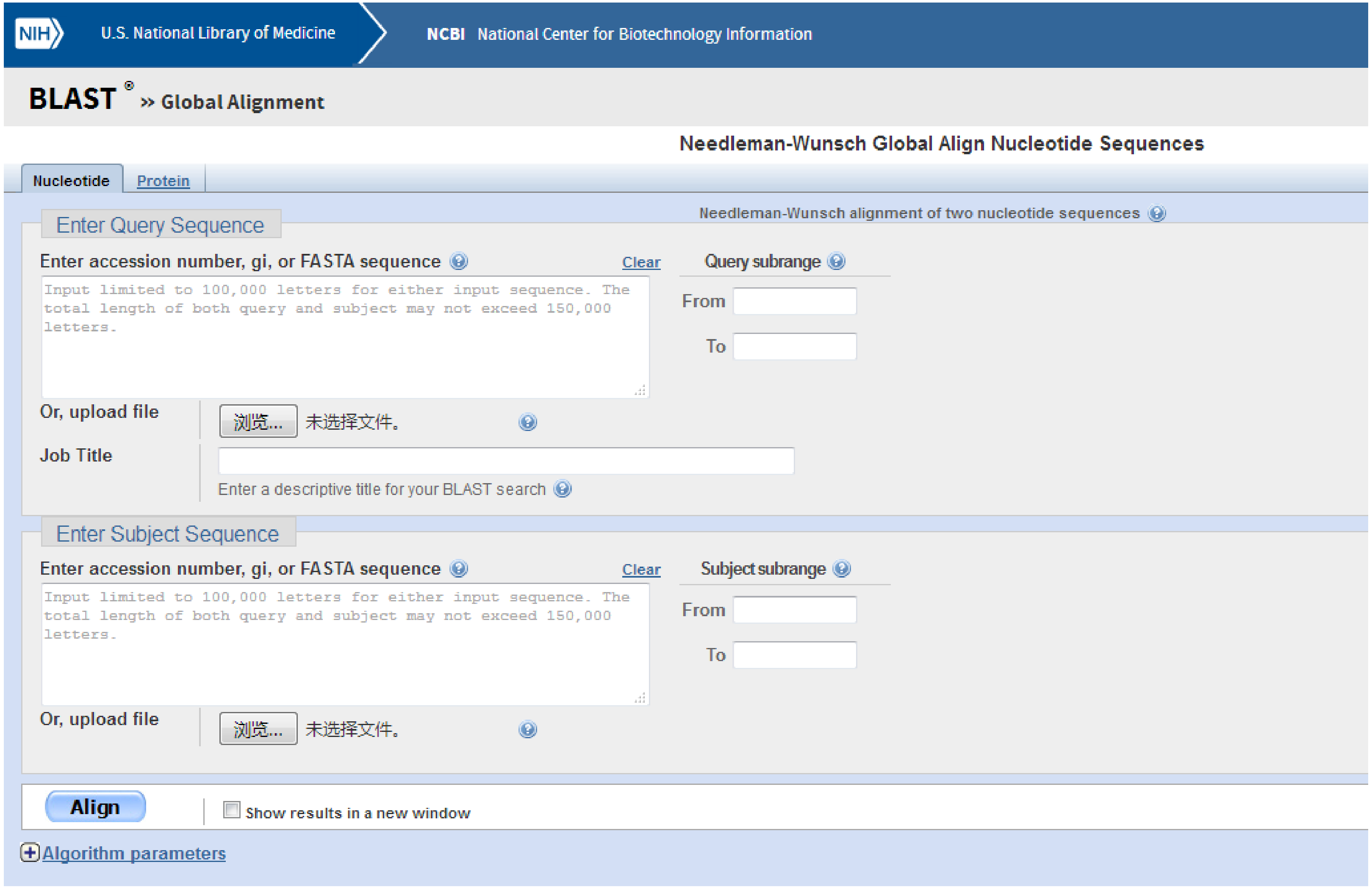
Last Update: 2013-10-28

Description

Design PCR primers from DNA sequence. Widely used (190k Google hits for "primer3"). From mispriming libraries to sequence quality data to the generation of internal oligos, primer3 does it. C&perl. Developers/testers/documenters needed.

Sequences alignment (two sequence)

NCBI homepage -> Blast homepage -> specialized searches -> Global Align



U.S. National Library of Medicine

NCBI National Center for Biotechnology Information

BLAST ® » Global Alignment

Needleman-Wunsch Global Align Nucleotide Sequences

Nucleotide | Protein

Needleman-Wunsch alignment of two nucleotide sequences

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear

Input limited to 100,000 letters for either input sequence. The total length of both query and subject may not exceed 150,000 letters.

Query subrange

From

To

Or, upload file 浏览... 未选择文件。

Job Title

Enter a descriptive title for your BLAST search

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence Clear

Input limited to 100,000 letters for either input sequence. The total length of both query and subject may not exceed 150,000 letters.

Subject subrange

From

To

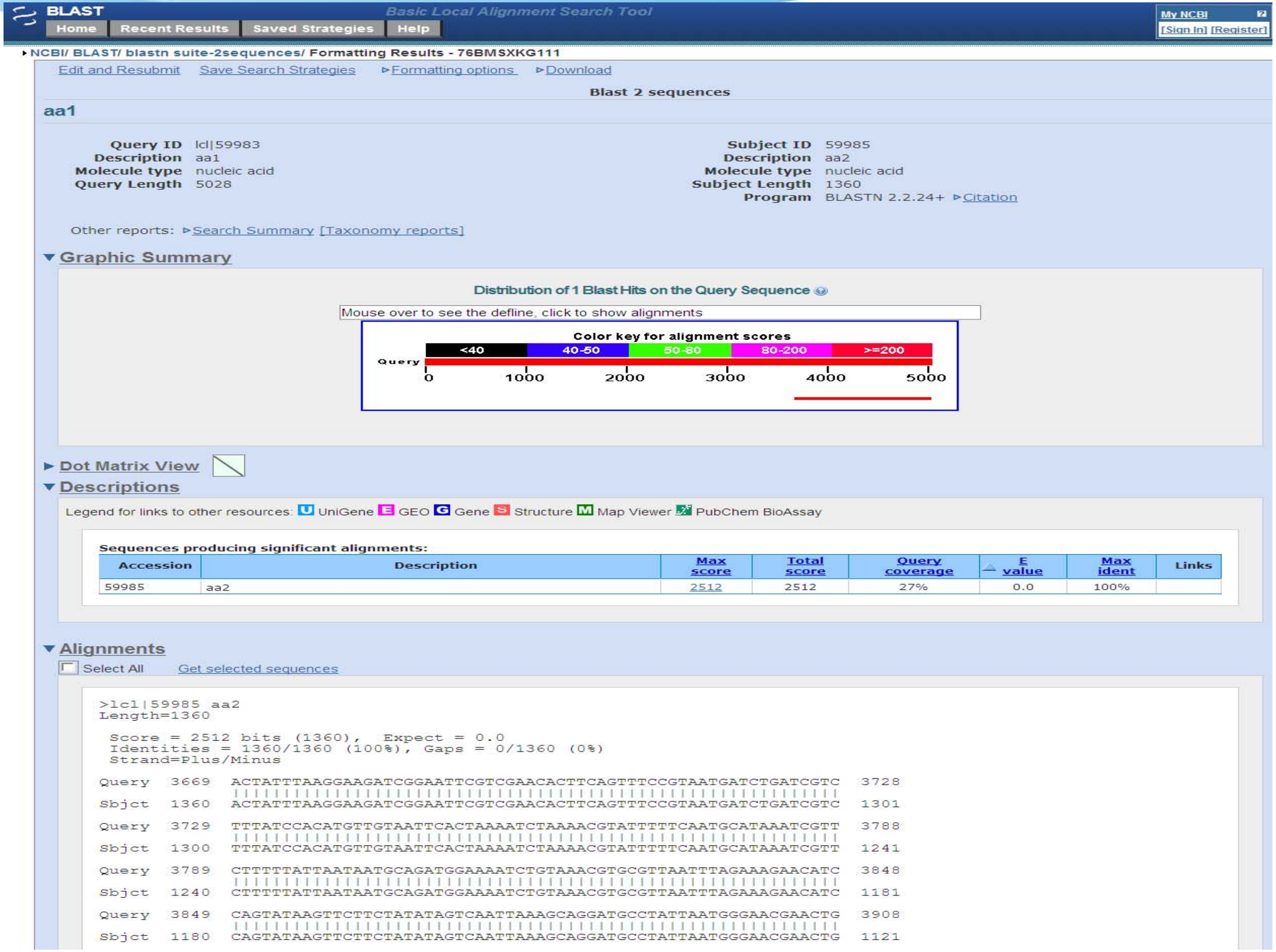
Or, upload file 浏览... 未选择文件。

Align

□ Show results in a new window

Algorithm parameters

Sequences alignment (two sequence)



NCBI/ BLAST/ blastn suite-2sequences/ Formatting Results - 76BMSXKG111

Edit and Resubmit   Save Search Strategies   ▹Formatting options   ▹Download

Blast 2 sequences

aa1

Query ID	lcl|59983	**Subject ID**	59985
Description	aa1	**Description**	aa2
Molecule type	nucleic acid	**Molecule type**	nucleic acid
Query Length	5028	**Subject Length**	1360
		Program	BLASTN 2.2.24+ ▹Citation

Other reports: ▹Search Summary [Taxonomy reports]

▾ Graphic Summary

Distribution of 1 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

▸ Dot Matrix View

▾ 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	Max ident	Links
59985	aa2	2512	2512	27%	0.0	100%	

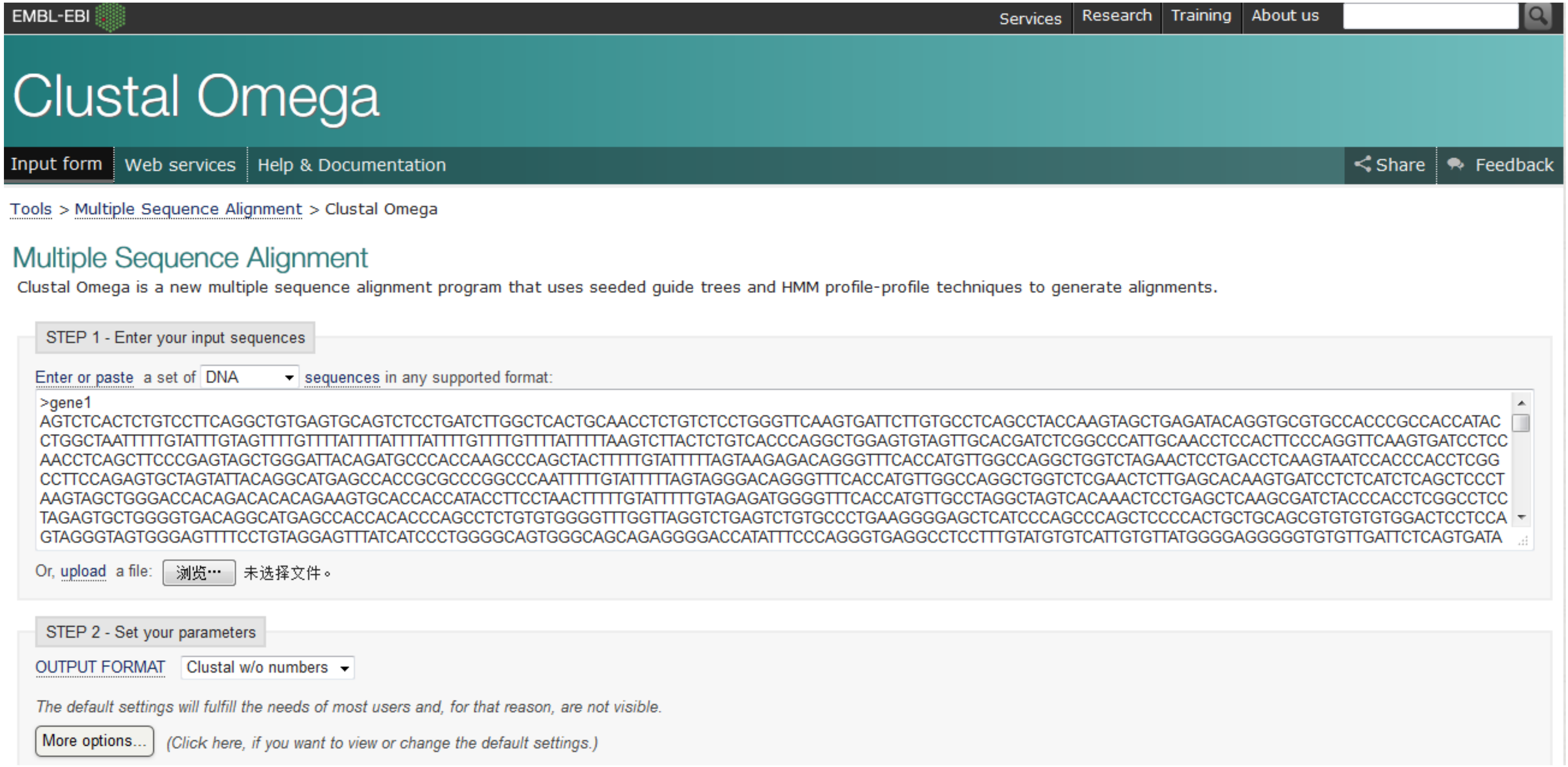
▾ Alignments

□ Select All   Get selected sequences

```
>lcl|59985 aa2
Length=1360

 Score = 2512 bits (1360),  Expect = 0.0
 Identities = 1360/1360 (100%), Gaps = 0/1360 (0%)
 Strand=Plus/Minus

Query  3669  ACTATTTAAGGAAGATCGGAATTCGTCGAACACTTCAGTTTCCGTAATGATCTGATCGTC  3728
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1360  ACTATTTAAGGAAGATCGGAATTCGTCGAACACTTCAGTTTCCGTAATGATCTGATCGTC  1301

Query  3729  TTTATCCACATGTTGTAATTCACTAAAATCTAAAACGTATTTTTCAATGCATAAATCGTT  3788
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1300  TTTATCCACATGTTGTAATTCACTAAAATCTAAAACGTATTTTTCAATGCATAAATCGTT  1241

Query  3789  CTTTTTATTAATAATGCAGATGGAAAATCTGTAAACGTGCGTTAATTTAGAAAGAACATC  3848
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1240  CTTTTTATTAATAATGCAGATGGAAAATCTGTAAACGTGCGTTAATTTAGAAAGAACATC  1181

Query  3849  CAGTATAAGTTCTTCTATATAGTCAATTAAAGCAGGATGCCTATTAATGGGAACGAACTG  3908
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1180  CAGTATAAGTTCTTCTATATAGTCAATTAAAGCAGGATGCCTATTAATGGGAACGAACTG  1121
```

Sequences comparison (multi-sequences)

http://www.ebi.ac.uk/Tools/msa/clustalo/



Tools > Multiple Sequence Alignment > Clustal Omega

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments.

STEP 1 - Enter your input sequences

Enter or paste a set of DNA sequences in any supported format:

```
>gene1
AGTCTCACTCTGTCCTTCAGGCTGTGAGTGCAGTCTCCTGATCTTGGCTCACTGCAACCTCTGTCTCCTGGGTTCAAGTGATTCTTGTGCCTCAGCCTACCAAGTAGCTGAGATACAGGTGCGTGCCACCCGCCACCATAC
CTGGCTAATTTTTGTATTTGTAGTTTTGTTTTATTTTATTTTATTTTGTTTTGTTTTATTTTTAAGTCTTACTCTGTCACCCAGGCTGGAGTGTAGTTGCACGATCTCGGCCCATTGCAACCTCCACTTCCCAGGTTCAAGTGATCCTCC
AACCTCAGCTTCCCGAGTAGCTGGGATTACAGATGCCCACCAAGCCCAGCTACTTTTTGTATTTTTAGTAAGAGACAGGGTTTCACCATGTTGGCCAGGCTGGTCTAGAACTCCTGACCTCAAGTAATCCACCCACCTCGG
CCTTCCAGAGTGCTAGTATTACAGGCATGAGCCACCGCGCCCGGCCCAATTTTTGTATTTTTAGTAGGGACAGGGTTTCACCATGTTGGCCAGGCTGGTCTCGAACTCTTGAGCACAAGTGATCCTCTCATCTCAGCTCCCT
AAGTAGCTGGGACCACAGACACACAGAAGTGCACCACCATACCTTCCTAACTTTTTGTATTTTTGTAGAGATGGGGTTTCACCATGTTGCCTAGGCTAGTCACAAACTCCTGAGCTCAAGCGATCTACCCACCTCGGCCTCC
TAGAGTGCTGGGGTGACAGGCATGAGCCACCACACCCAGCCTCTGTGTGGGGTTTGGTTAGGTCTGAGTCTGTGCCCTGAAGGGGAGCTCATCCCAGCCCAGCTCCCCACTGCTGCAGCGTGTGTGTGGACTCCTCCA
GTAGGGTAGTGGGAGTTTTCCTGTAGGAGTTTATCATCCCTGGGGCAGTGGGCAGCAGAGGGGACCATATTTCCCAGGGTGAGGCCTCCTTTGTATGTGTCATTGTGTTATGGGGAGGGGGTGTGTTGATTCTCAGTGATA
```

Or, upload a file: 浏览… 未选择文件。

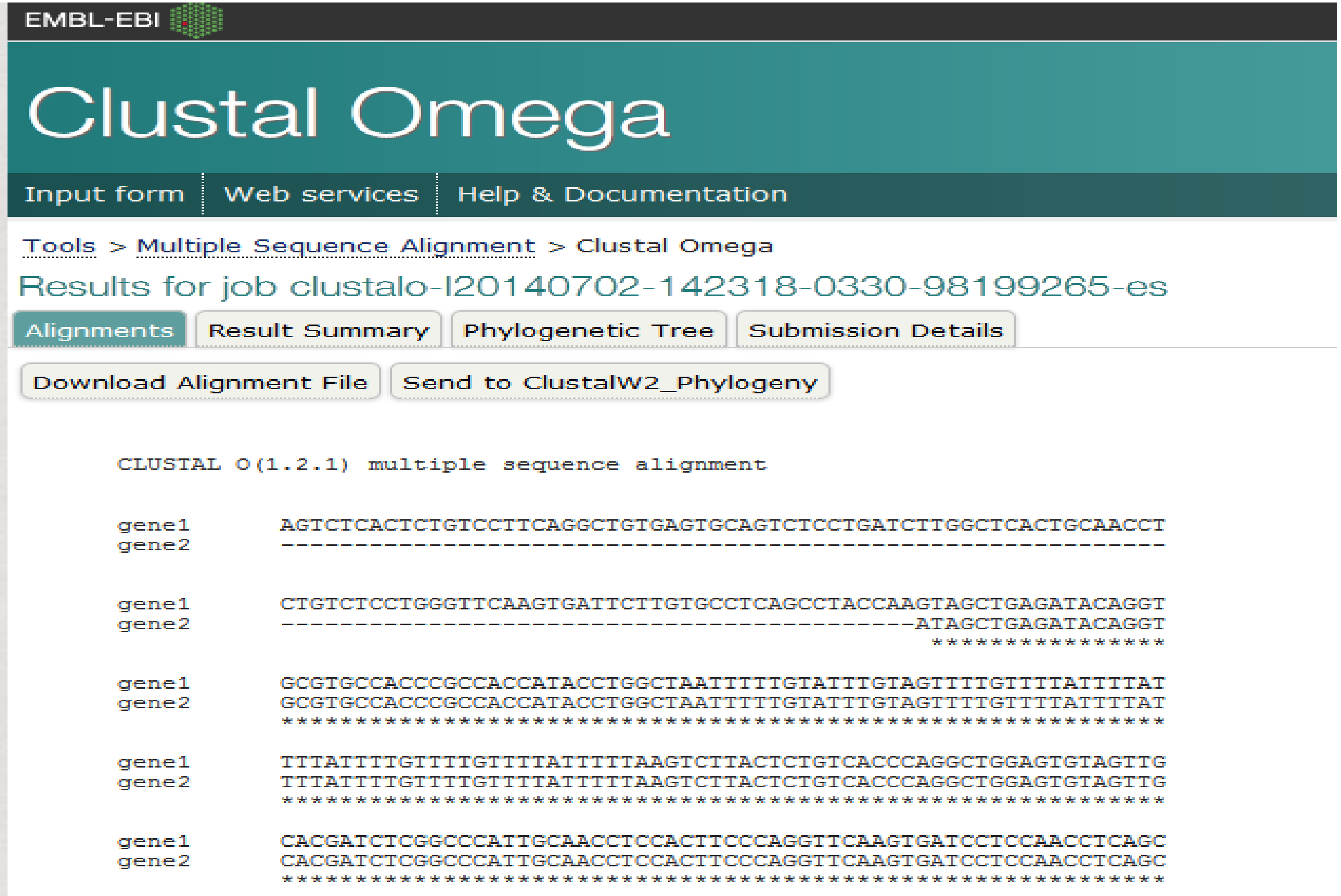
STEP 2 - Set your parameters

OUTPUT FORMAT: Clustal w/o numbers

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(Click here, if you want to view or change the default settings.)*

Sequences comparison (multi-sequences)



EMBL-EBI

Clustal Omega

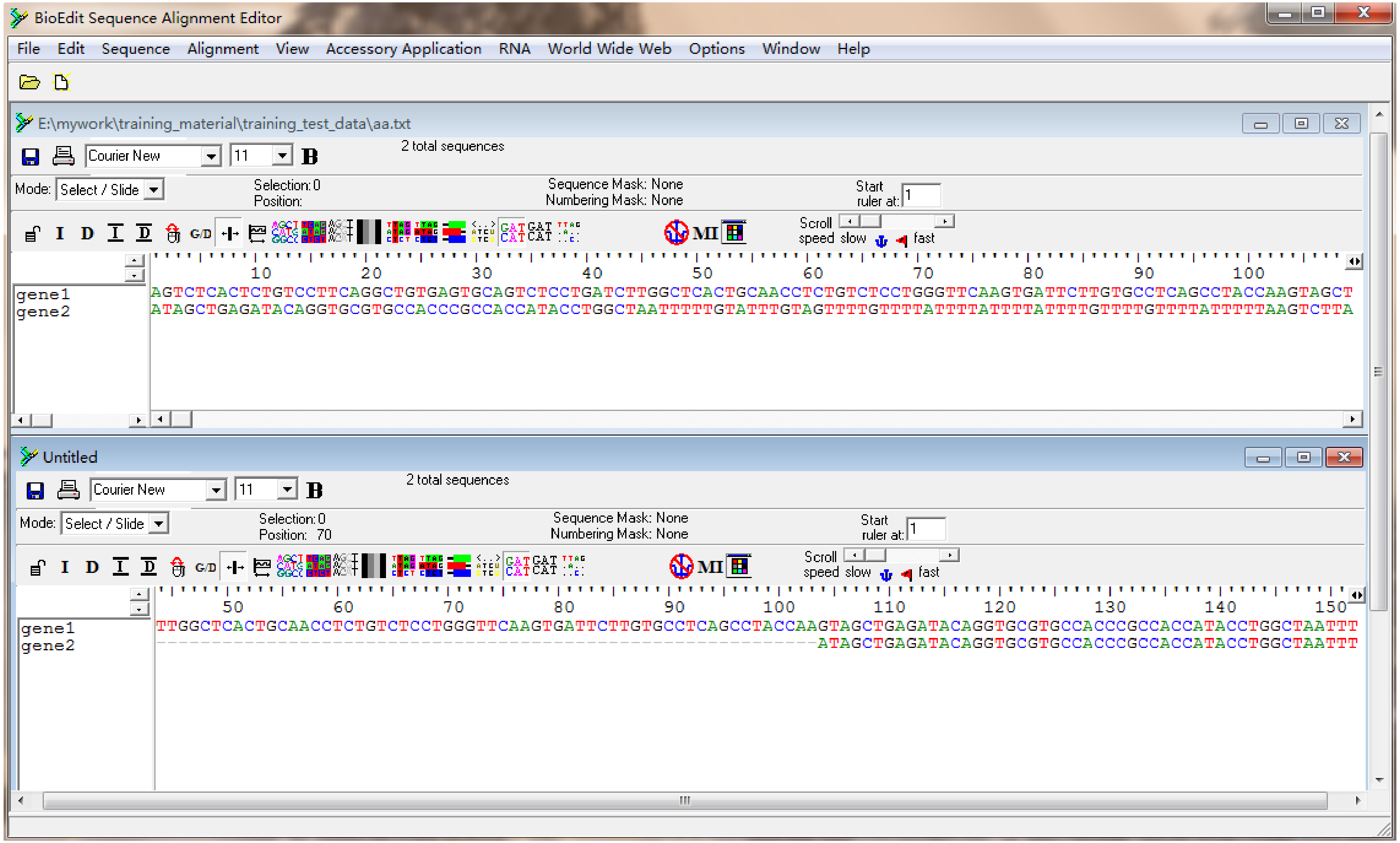
Results for job clustalo-I20140702-142318-0330-98199265-es

```
CLUSTAL O(1.2.1) multiple sequence alignment


gene1           AGTCTCACTCTGTCCTTCAGGCTGTGAGTGCAGTCTCCTGATCTTGGCTCACTGCAACCT
gene2           ------------------------------------------------------------
                                                                            

gene1           CTGTCTCCTGGGTTCAAGTGATTCTTGTGCCTCAGCCTACCAAGTAGCTGAGATACAGGT
gene2           -------------------------------------------ATAGCTGAGATACAGGT
                                                            ****************

gene1           GCGTGCCACCCGCCACCATACCTGGCTAATTTTTGTATTTGTAGTTTTGTTTTATTTTAT
gene2           GCGTGCCACCCGCCACCATACCTGGCTAATTTTTGTATTTGTAGTTTTGTTTTATTTTAT
                ************************************************************

gene1           TTTATTTTGTTTTGTTTTATTTTTAAGTCTTACTCTGTCACCCAGGCTGGAGTGTAGTTG
gene2           TTTATTTTGTTTTGTTTTATTTTTAAGTCTTACTCTGTCACCCAGGCTGGAGTGTAGTTG
                ************************************************************

gene1           CACGATCTCGGCCCATTGCAACCTCCACTTCCCAGGTTCAAGTGATCCTCCAACCTCAGC
gene2           CACGATCTCGGCCCATTGCAACCTCCACTTCCCAGGTTCAAGTGATCCTCCAACCTCAGC
                ************************************************************
```

Sequences comparison (multi-sequences)

Software: BioEdit

File -> open -> sequence -> Accessory Application -> clustalw

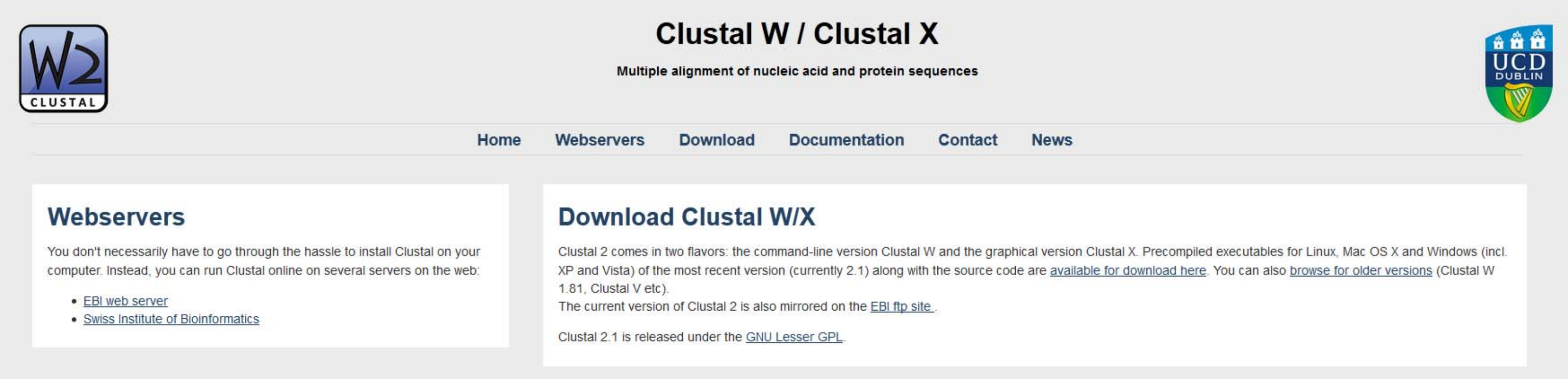

BioEdit Sequence Alignment Editor
File Edit Sequence Alignment View Accessory Application RNA World Wide Web Options Window Help
E:\mywork\training_material\training_test_data\aa.txt
Courier New
11
2 total sequences
Mode: Select / Slide
Selection: 0
Position:
Sequence Mask: None
Numbering Mask: None
Start ruler at: 1
Scroll speed slow fast
10 20 30 40 50 60 70 80 90 100

```
gene1  AGTCTCACTCTGTCCTTCAGGCTGTGAGTGCAGTCTCCTGATCTTGGCTCACTGCAACCTCTGTCTCCTGGGTTCAAGTGATTCTTGTGCCTCAGCCTACCAAGTAGCT
gene2  ATAGCTGAGATACAGGTGCGTGCCACCCGCCACCATACCTGGCTAATTTTTGTATTTGTAGTTTTGTTTTATTTTATTTTATTTTGTTTTGTTTTATTTTTAAGTCTTA
```

```
gene1  TTGGCTCACTGCAACCTCTGTCTCCTGGGTTCAAGTGATTCTTGTGCCTCAGCCTACCAAGTAGCTGAGATACAGGTGCGTGCCACCCGCCACCATACCTGGCTAATTT
gene2  ------------------------------------------------------------ATAGCTGAGATACAGGTGCGTGCCACCCGCCACCATACCTGGCTAATTT
```

Sequences comparison (multi-sequences)

http://www.clustal.org/clustal2/



Clustal W / Clustal X

Multiple alignment of nucleic acid and protein sequences

Webservers

You don't necessarily have to go through the hassle to install Clustal on your computer. Instead, you can run Clustal online on several servers on the web:

- EBI web server
- Swiss Institute of Bioinformatics

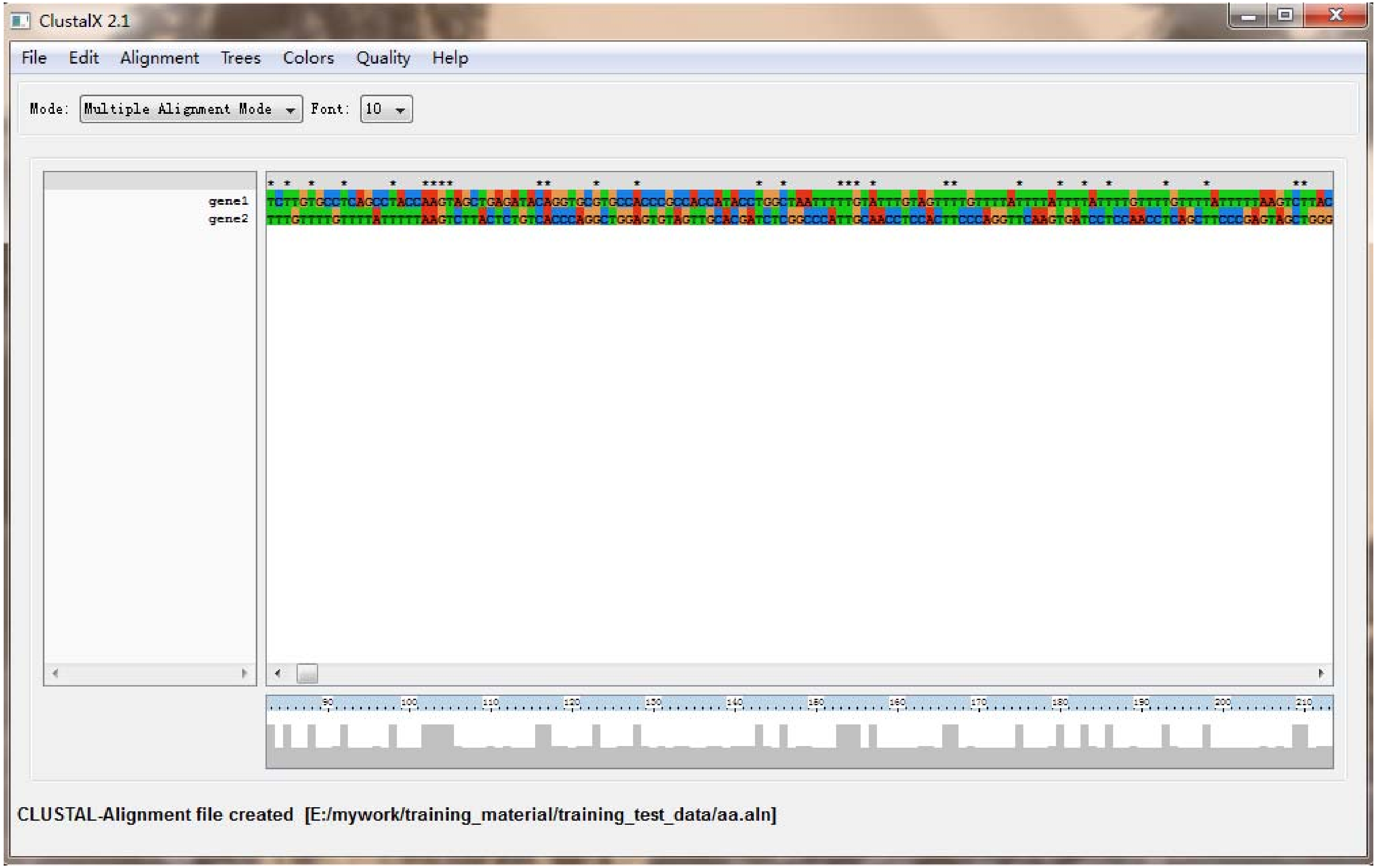
Download Clustal W/X

Clustal 2 comes in two flavors: the command-line version Clustal W and the graphical version Clustal X. Precompiled executables for Linux, Mac OS X and Windows (incl. XP and Vista) of the most recent version (currently 2.1) along with the source code are available for download here. You can also browse for older versions (Clustal W 1.81, Clustal V etc).
The current version of Clustal 2 is also mirrored on the EBI ftp site .

Clustal 2.1 is released under the GNU Lesser GPL.

Sequences comparison (multi-sequences)

Software: clustalw / clustalX   menu : load sequence



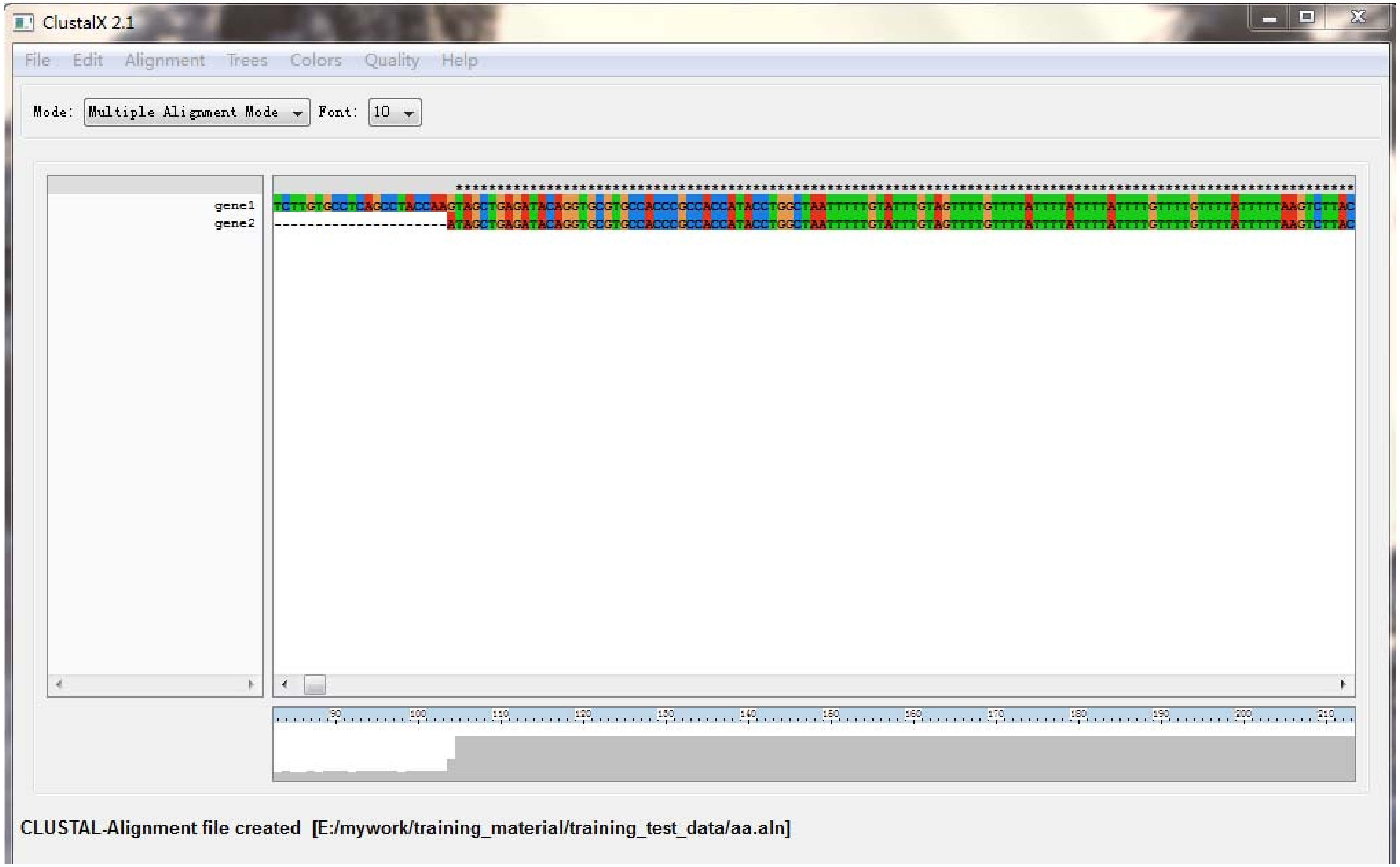
Mode: Multiple Alignment Mode   Font: 10

```
gene1 TCTTGTGCCTCAGCCTACCAAGTAGCTGAGATACAGGTGCGTGCCACCCGCCACCATACCTGGCTAATTTTTGTATTTGTAGTTTTGTTTTATTTTATTTTATTTTGTTTTGTTTTATTTTTAAGTCTTAC
gene2 TTTGTTTTGTTTTATTTTTAAGTCTTACTCTGTCACCCAGGCTGGAGTGTAGTTGCACGATCTCGGCCCATTGCAACCTCCACTTCCCAGGTTCAAGTGATCCTCCAACCTCAGCTTCCCGAGTAGCTGGG
```

CLUSTAL-Alignment file created [E:/mywork/training_material/training_test_data/aa.aln]

Sequences comparison (multi-sequences)

Software: clustalw / clustalX menu : Alignment -> do complete alignment



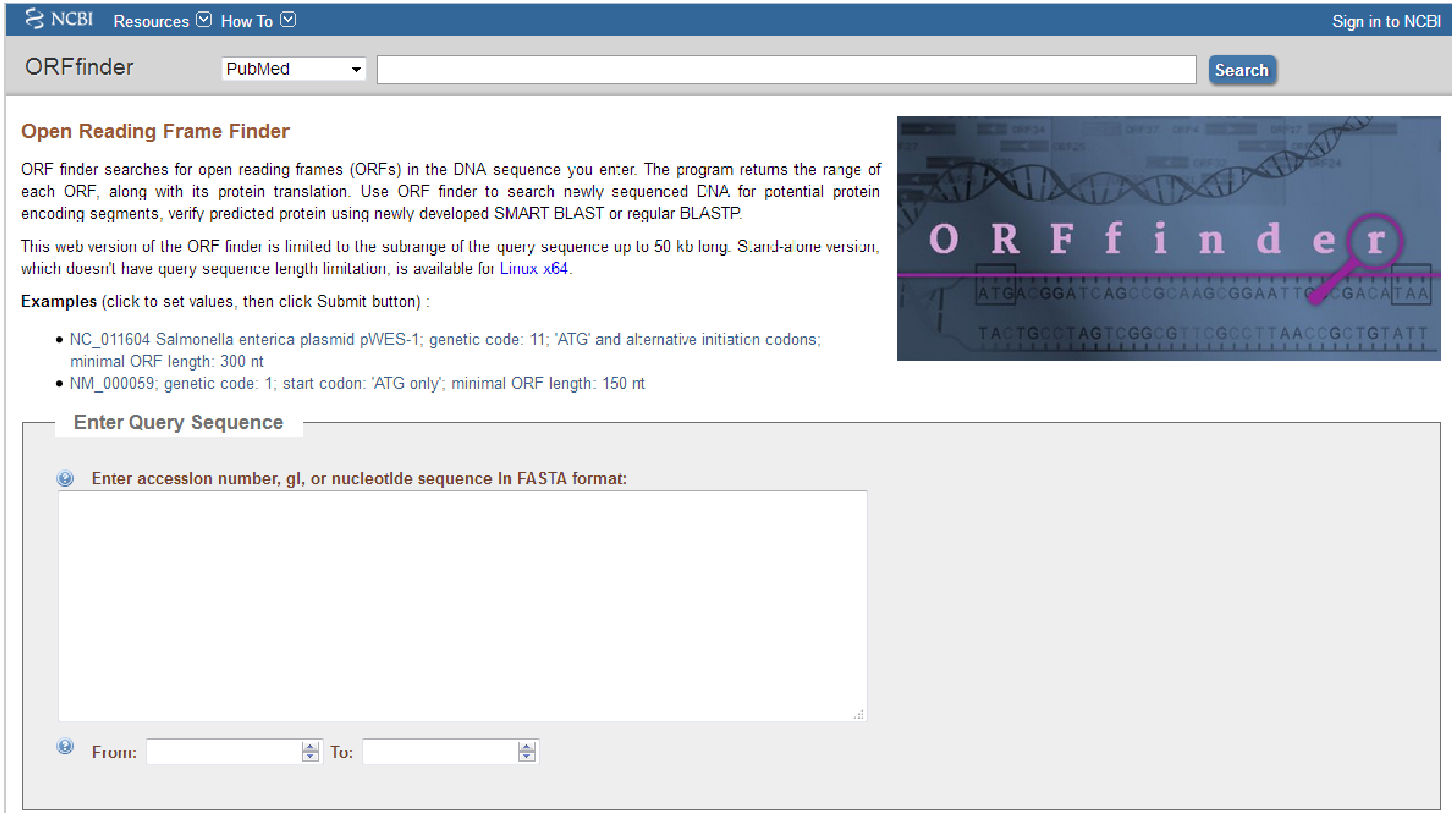
Mode: Multiple Alignment Mode Font: 10

```
                             *************************************************************************************************************
gene1 TCTTGTGCCTCAGCCTACCAAGTAGCTGAGATACAGGTGCGTGCCACCCGCCACCATACCTGGCTAATTTTTGTATTTGTAGTTTTGTTTTATTTTATTTTATTTTGTTTTGTTTTATTTTTAAGTCTTAC
gene2 ---------------------ATAGCTGAGATACAGGTGCGTGCCACCCGCCACCATACCTGGCTAATTTTTGTATTTGTAGTTTTGTTTTATTTTATTTTATTTTGTTTTGTTTTATTTTTAAGTCTTAC
```

CLUSTAL-Alignment file created [E:/mywork/training_material/training_test_data/aa.aln]

Analysis of open reading frame

https://www.ncbi.nlm.nih.gov/orffinder/



Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.

This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for Linux x64.

Examples (click to set values, then click Submit button) :

- NC_011604 Salmonella enterica plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- NM_000059; genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt

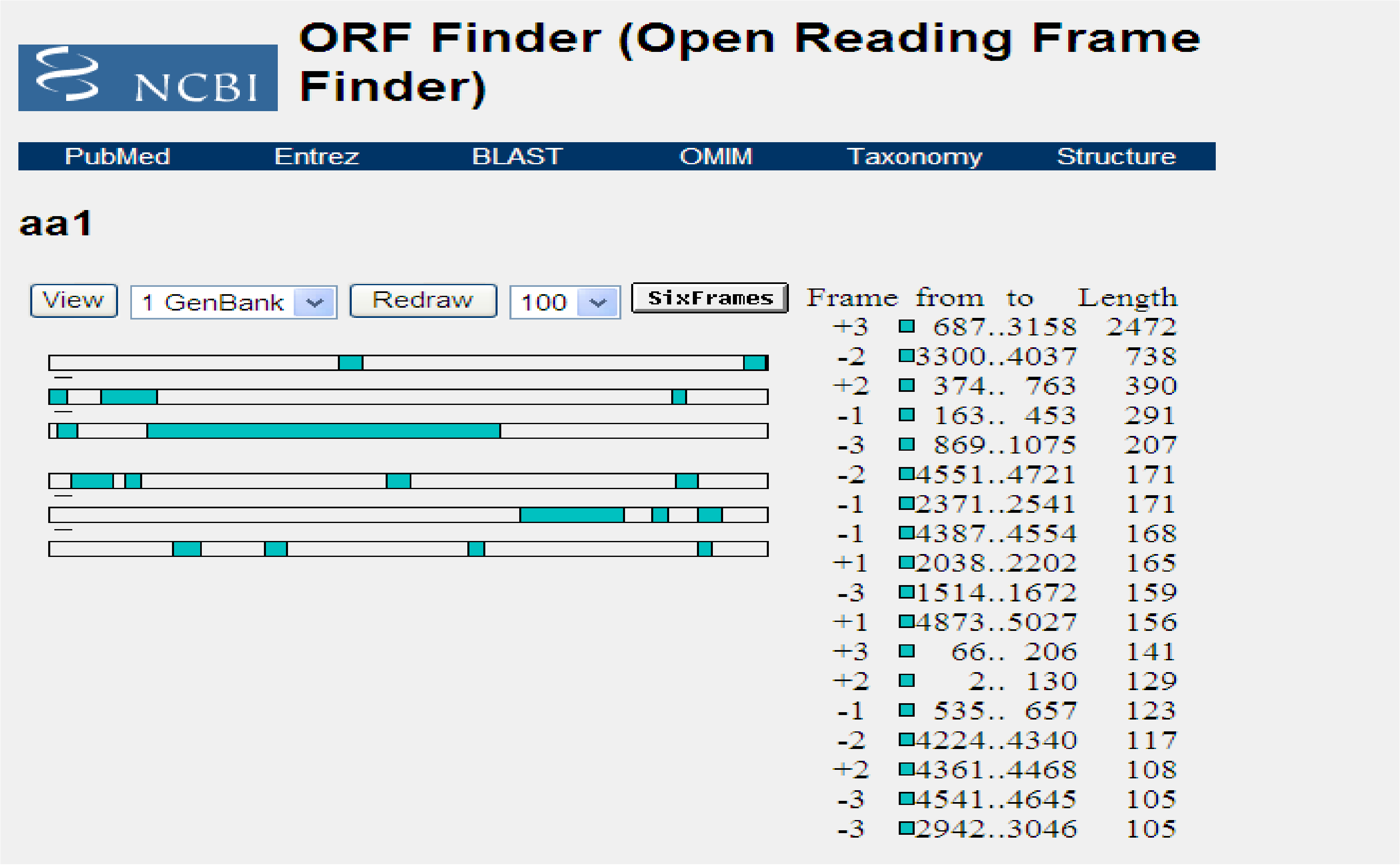
ORFfinder
ATGACGGATCAGCCGCAAGCGGAATTCGACATAA
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT

Enter Query Sequence

Enter accession number, gi, or nucleotide sequence in FASTA format:

From: To:

Analysis of open reading frame



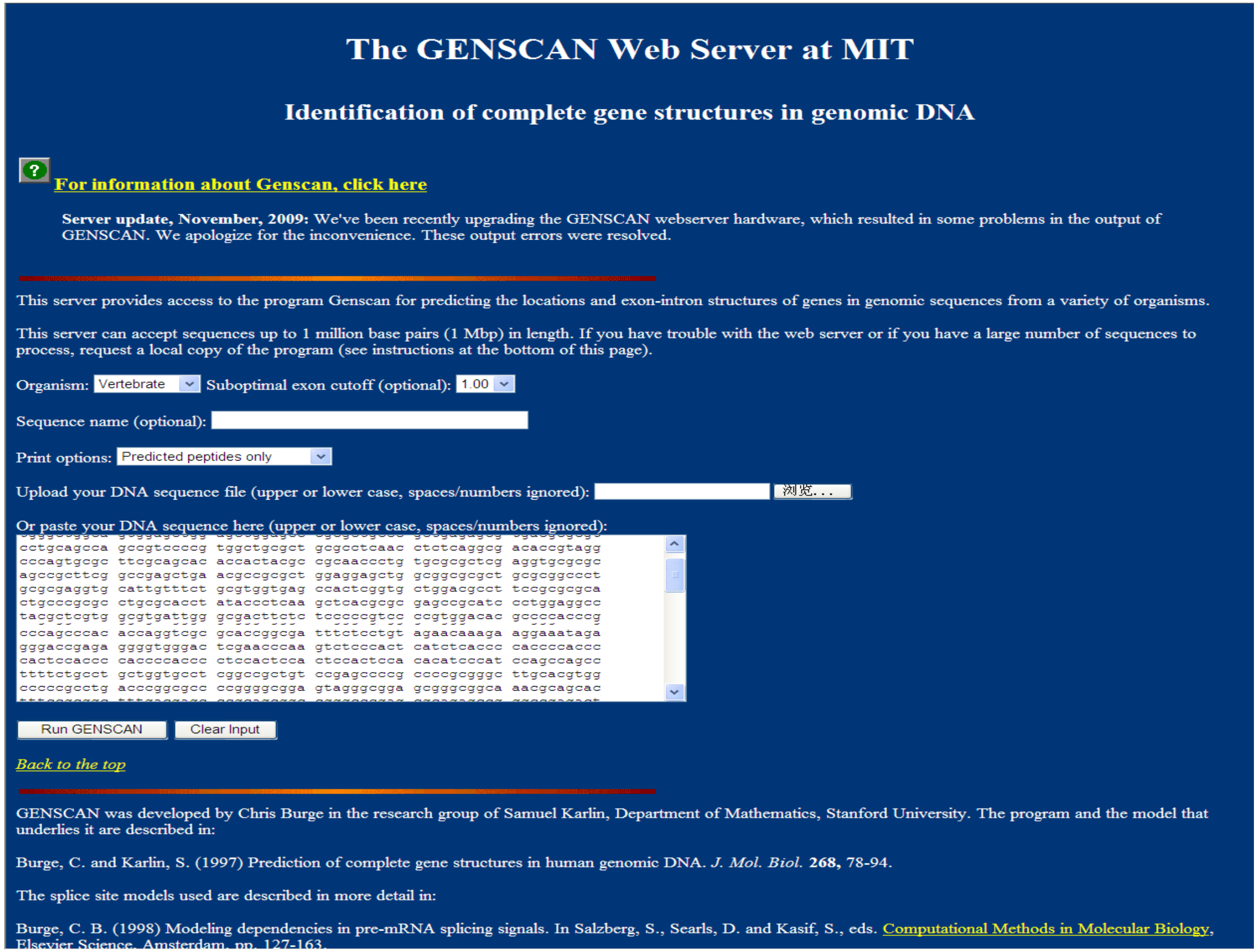
ORF Finder (Open Reading Frame Finder)

aa1

Frame		from	to	Length
+3	■	687	3158	2472
-2	■	3300	4037	738
+2	■	374	763	390
-1	■	163	453	291
-3	■	869	1075	207
-2	■	4551	4721	171
-1	■	2371	2541	171
-1	■	4387	4554	168
+1	■	2038	2202	165
-3	■	1514	1672	159
+1	■	4873	5027	156
+3	■	66	206	141
+2	■	2	130	129
-1	■	535	657	123
-2	■	4224	4340	117
+2	■	4361	4468	108
-3	■	4541	4645	105
-3	■	2942	3046	105

Gene prediction

http://genes.mit.edu/GENSCAN.html



The GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA

? For information about Genscan, click here

Server update, November, 2009: We've been recently upgrading the GENSCAN webserver hardware, which resulted in some problems in the output of GENSCAN. We apologize for the inconvenience. These output errors were resolved.

This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page).

Organism: Vertebrate Suboptimal exon cutoff (optional): 1.00

Sequence name (optional):

Print options: Predicted peptides only

Upload your DNA sequence file (upper or lower case, spaces/numbers ignored):

Or paste your DNA sequence here (upper or lower case, spaces/numbers ignored):

```
cctgcagcca gccgtccccg tggctgcgct gcgcctcaac ctctcaggcg acaccgtagg
cccagtgcgc ttcgcagcac accactacgc cgcaaccctg tgcgcgctcg aggtgcgcgc
agccgcttcg gccgagctga acgccgcgct ggaggagctg gcggcgcgct gcgcggccct
gcgcgaggtg cattgtttct gcgtggtgag ccactcggtg ctggacgcct tccgcgcgca
ctgcccgcgc ctgcgcacct ataccctcaa gctcacgcgc gagccgcatc cctggaggcc
tacgctcgtg gcgtgattgg gcgacttctc tccccgtcc ccgtggacac gccccacccg
cccagcccac accaggtcgc gcaccggcga tttctcctgt agaacaaaga aggaaataga
gggaccgaga ggggtgggac tcgaacccaa gtctcccact catctcaccc caccccaccc
cactccaccc caccccaccc ctccactcca ctccactcca cacatcccat ccagccagcc
ttttctgcct gctggtgcct cggccgctgt ccgagccccg ccccgcgggc ttgcacgtgg
cccccgcctg acccggcgcc ccggggcgga gtagggcgga gcgggcggca aacgcagcac
```

Back to the top

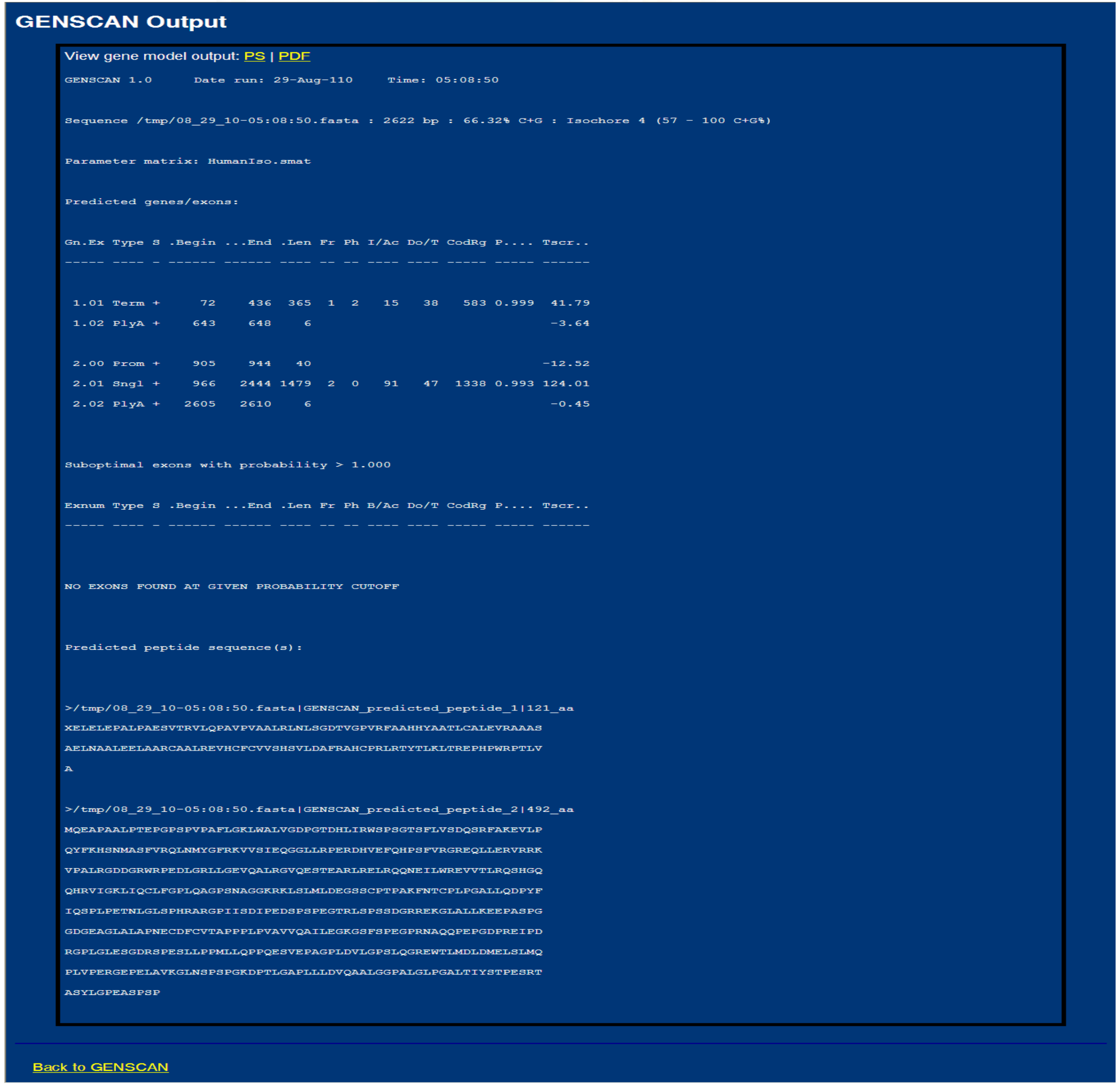
GENSCAN was developed by Chris Burge in the research group of Samuel Karlin, Department of Mathematics, Stanford University. The program and the model that underlies it are described in:

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78-94.

The splice site models used are described in more detail in:

Burge, C. B. (1998) Modeling dependencies in pre-mRNA splicing signals. In Salzberg, S., Searls, D. and Kasif, S., eds. Computational Methods in Molecular Biology, Elsevier Science, Amsterdam, pp. 127-163.

Gene prediction



GENSCAN Output

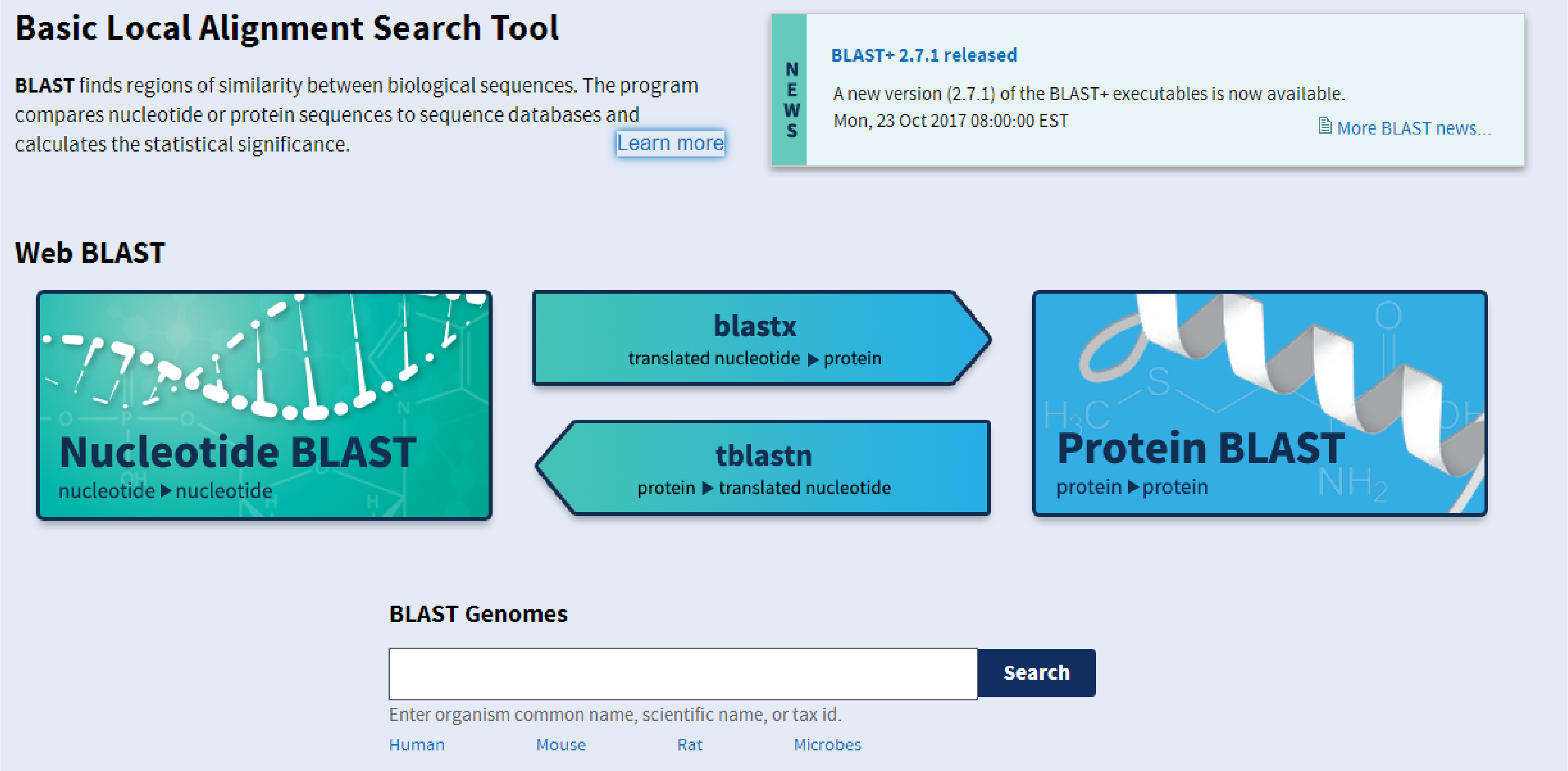
View gene model output: PS | PDF

```
GENSCAN 1.0      Date run: 29-Aug-110     Time: 05:08:50

Sequence /tmp/08_29_10-05:08:50.fasta : 2622 bp : 66.32% C+G : Isochore 4 (57 - 100 C+G%)

Parameter matrix: HumanIso.smat

Predicted genes/exons:

Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------

 1.01 Term +     72    436  365  1  2   15   38   583 0.999  41.79
 1.02 PlyA +    643    648    6                               -3.64

 2.00 Prom +    905    944   40                              -12.52
 2.01 Sngl +    966   2444 1479  2  0   91   47  1338 0.993 124.01
 2.02 PlyA +   2605   2610    6                               -0.45


Suboptimal exons with probability > 1.000

Exnum Type S .Begin ...End .Len Fr Ph B/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------



NO EXONS FOUND AT GIVEN PROBABILITY CUTOFF


Predicted peptide sequence(s):


>/tmp/08_29_10-05:08:50.fasta|GENSCAN_predicted_peptide_1|121_aa
XELELEPALPAESVTRVLQPAVPVAALRLNLSGDTVGPVRFAAHHYAATLCALEVRAAAS
AELNAALEELAARCAALREVHCFCVVSHSVLDAFRAHCPRLRTYTLKLTREPHPWRPTLV
A

>/tmp/08_29_10-05:08:50.fasta|GENSCAN_predicted_peptide_2|492_aa
MQEAPAALPTEPGPSPVPAFLGKLWALVGDPGTDHLIRWSPSGTSFLVSDQSRFAKEVLP
QYFKHSNMASFVRQLNMYGFRKVVSIEQGGLLRPERDHVEFQHPSFVRGREQLLERVRRK
VPALRGDDGRWRPEDLGRLLGEVQALRGVQESTEARLRELRQQNEILWREVVTLRQSHGQ
QHRVIGKLIQCLFGPLQAGPSNAGGKRKLSLMLDEGSSCPTPAKFNTCPLPGALLQDPYF
IQSPLPETNLGLSPHRARGPIISDIPEDSPSPEGTRLSPSSDGRREKGLALLKEEPASPG
GDGEAGLALAPNECDFCVTAPPPLPVAVVQAILEGKGSFSPEGPRNAQQPEPGDPREIPD
RGPLGLESGDRSPESLLPPMLLQPPQESVEPAGPLDVLGPSLQGREWTLMDLDMELSLMQ
PLVPERGEPELAVKGLNSPSPGKDPTLGAPLLLDVQAALGGPALGLPGALTIYSTPESRT
ASYLGPEASPSP
```

Back to GENSCAN

Sequence localization in genome

NCBI homepage -> Blast homepage -> BLAST Genomes -> Human



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

BLAST+ 2.7.1 released

A new version (2.7.1) of the BLAST+ executables is now available.

Mon, 23 Oct 2017 08:00:00 EST

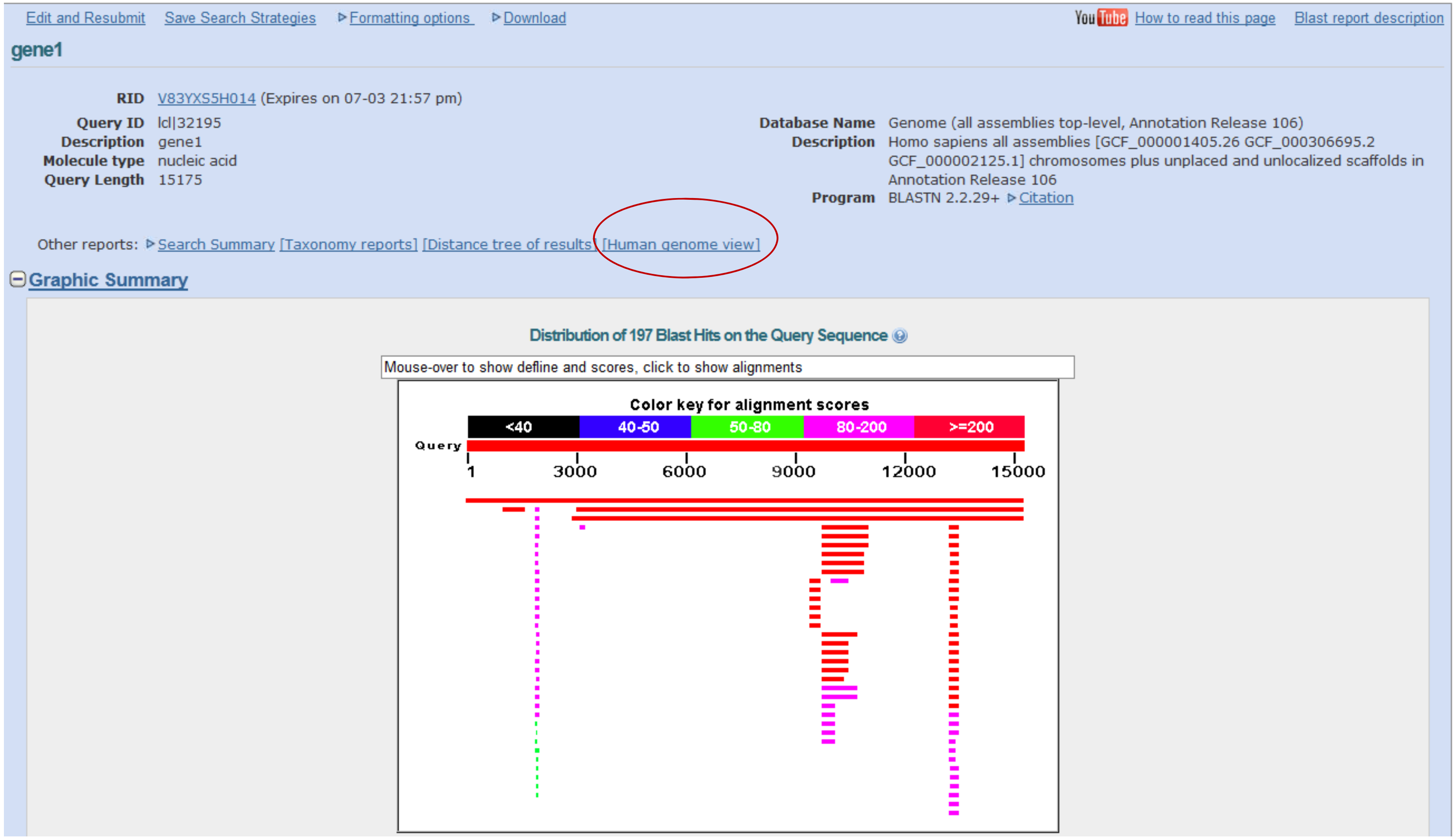
More BLAST news...

Web BLAST

BLAST Genomes

Enter organism common name, scientific name, or tax id.

Sequence localization in genome



Edit and Resubmit   Save Search Strategies   ▷Formatting options   ▷Download   YouTube How to read this page   Blast report description

gene1

RID V83YXS5H014 (Expires on 07-03 21:57 pm)

Query ID lcl|32195
Description gene1
Molecule type nucleic acid
Query Length 15175

Database Name Genome (all assemblies top-level, Annotation Release 106)
Description Homo sapiens all assemblies [GCF_000001405.26 GCF_000306695.2 GCF_000002125.1] chromosomes plus unplaced and unlocalized scaffolds in Annotation Release 106
Program BLASTN 2.2.29+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [Human genome view]

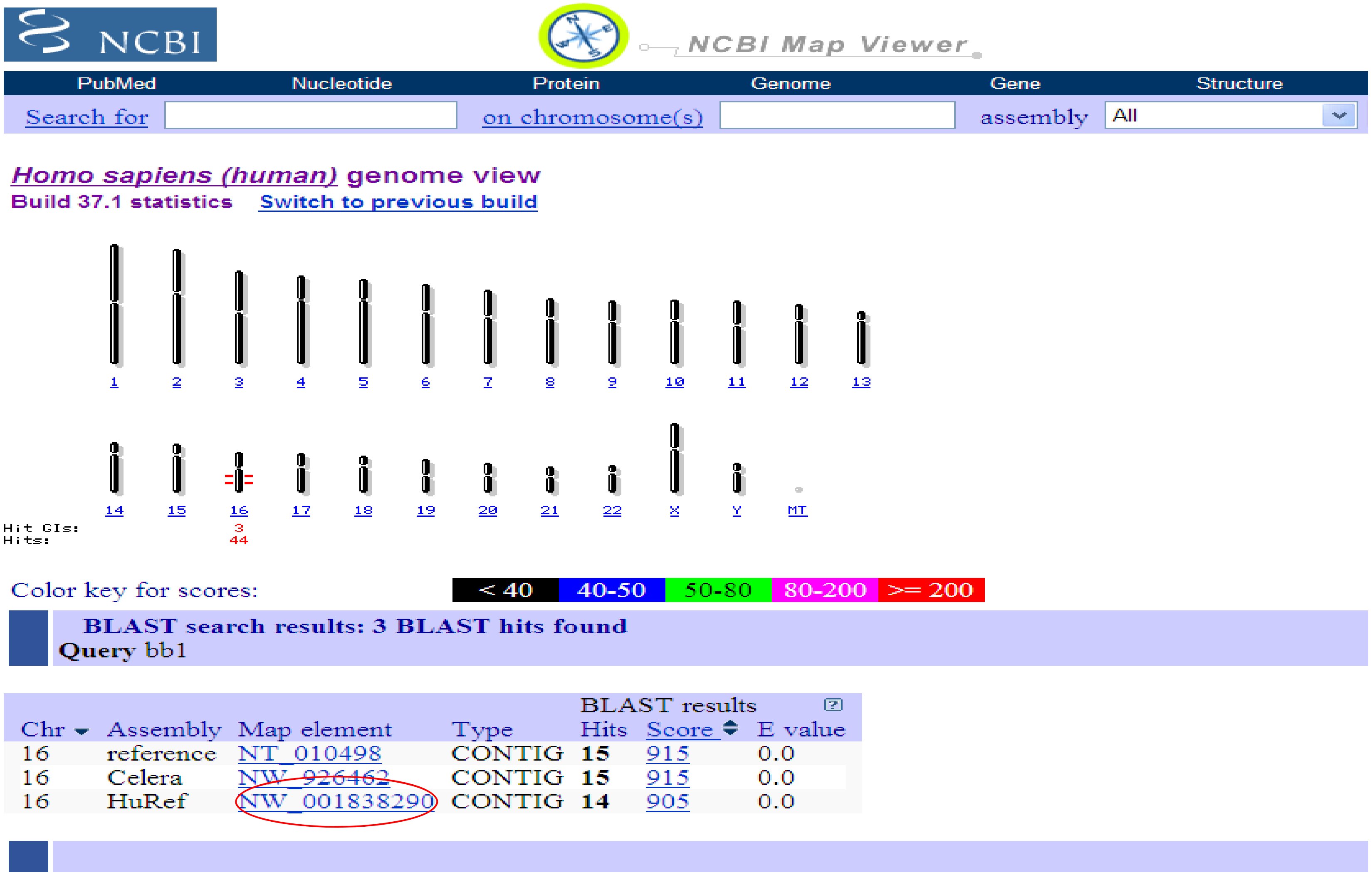
⊟ Graphic Summary

Distribution of 197 Blast Hits on the Query Sequence

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

Sequence localization in genome

Blast result -> view report -> human genome view



Homo sapiens (human) genome view

Build 37.1 statistics   Switch to previous build

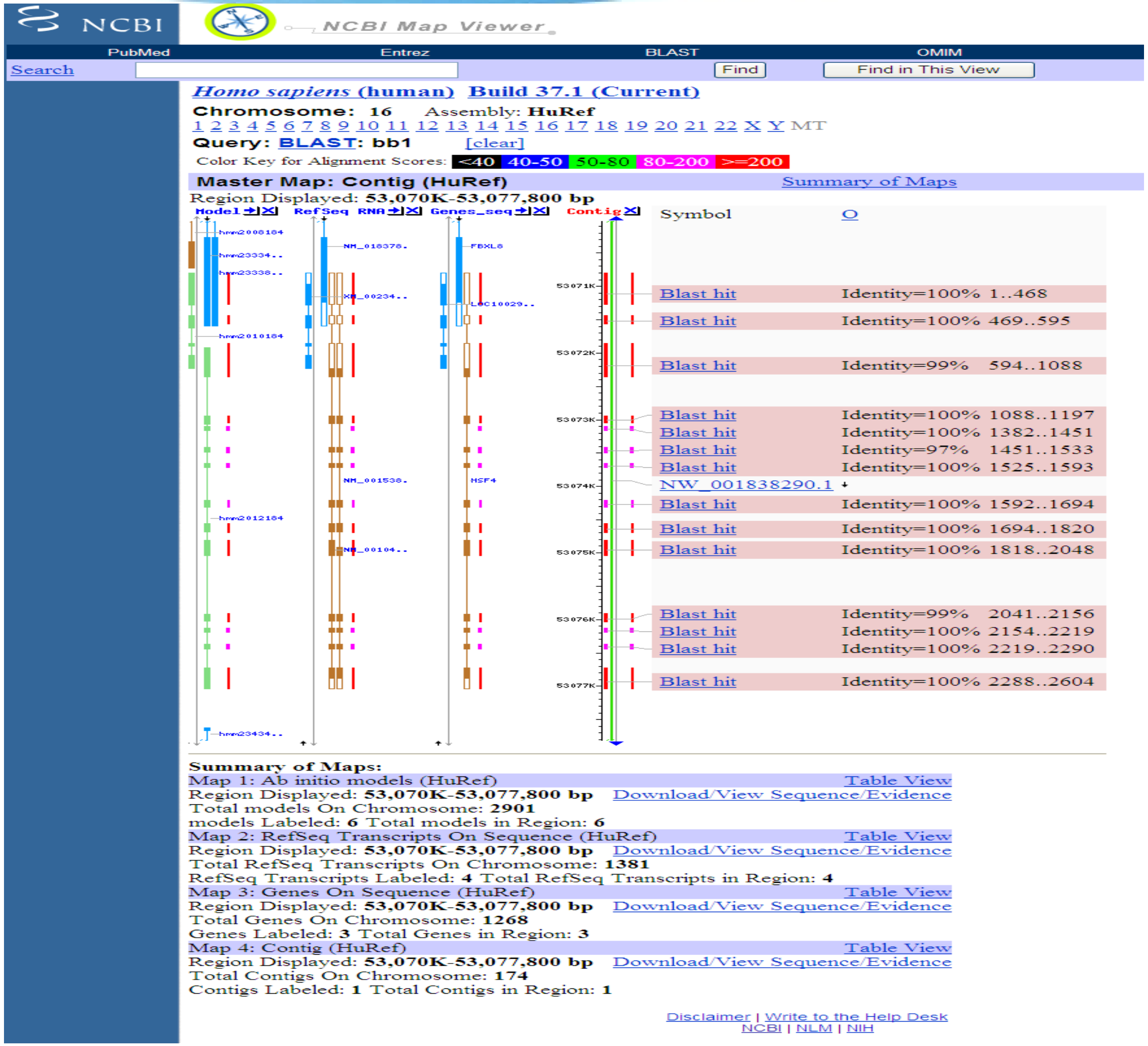
Color key for scores: < 40 | 40-50 | 50-80 | 80-200 | >= 200

BLAST search results: 3 BLAST hits found

Query bb1

Chr	Assembly	Map element	Type	BLAST results Hits	Score	E value
16	reference	NT_010498	CONTIG	**15**	915	0.0
16	Celera	NW_926462	CONTIG	**15**	915	0.0
16	HuRef	NW_001838290	CONTIG	**14**	905	0.0

Sequence localization in genome



Homo sapiens (human) Build 37.1 (Current)

Chromosome: 16 Assembly: **HuRef**

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT

Query: BLAST: bb1 [clear]

Color Key for Alignment Scores: <40 40-50 50-80 80-200 >=200

Master Map: Contig (HuRef) Summary of Maps

Region Displayed: **53,070K-53,077,800 bp**

Symbol	O
Blast hit	Identity=100% 1..468
Blast hit	Identity=100% 469..595
Blast hit	Identity=99% 594..1088
Blast hit	Identity=100% 1088..1197
Blast hit	Identity=100% 1382..1451
Blast hit	Identity=97% 1451..1533
Blast hit	Identity=100% 1525..1593
NW_001838290.1 +	
Blast hit	Identity=100% 1592..1694
Blast hit	Identity=100% 1694..1820
Blast hit	Identity=100% 1818..2048
Blast hit	Identity=99% 2041..2156
Blast hit	Identity=100% 2154..2219
Blast hit	Identity=100% 2219..2290
Blast hit	Identity=100% 2288..2604

Summary of Maps:

Map 1: Ab initio models (HuRef) Table View
Region Displayed: **53,070K-53,077,800 bp** Download/View Sequence/Evidence
Total models On Chromosome: **2901**
models Labeled: **6** Total models in Region: **6**

Map 2: RefSeq Transcripts On Sequence (HuRef) Table View
Region Displayed: **53,070K-53,077,800 bp** Download/View Sequence/Evidence
Total RefSeq Transcripts On Chromosome: **1381**
RefSeq Transcripts Labeled: **4** Total RefSeq Transcripts in Region: **4**

Map 3: Genes On Sequence (HuRef) Table View
Region Displayed: **53,070K-53,077,800 bp** Download/View Sequence/Evidence
Total Genes On Chromosome: **1268**
Genes Labeled: **3** Total Genes in Region: **3**

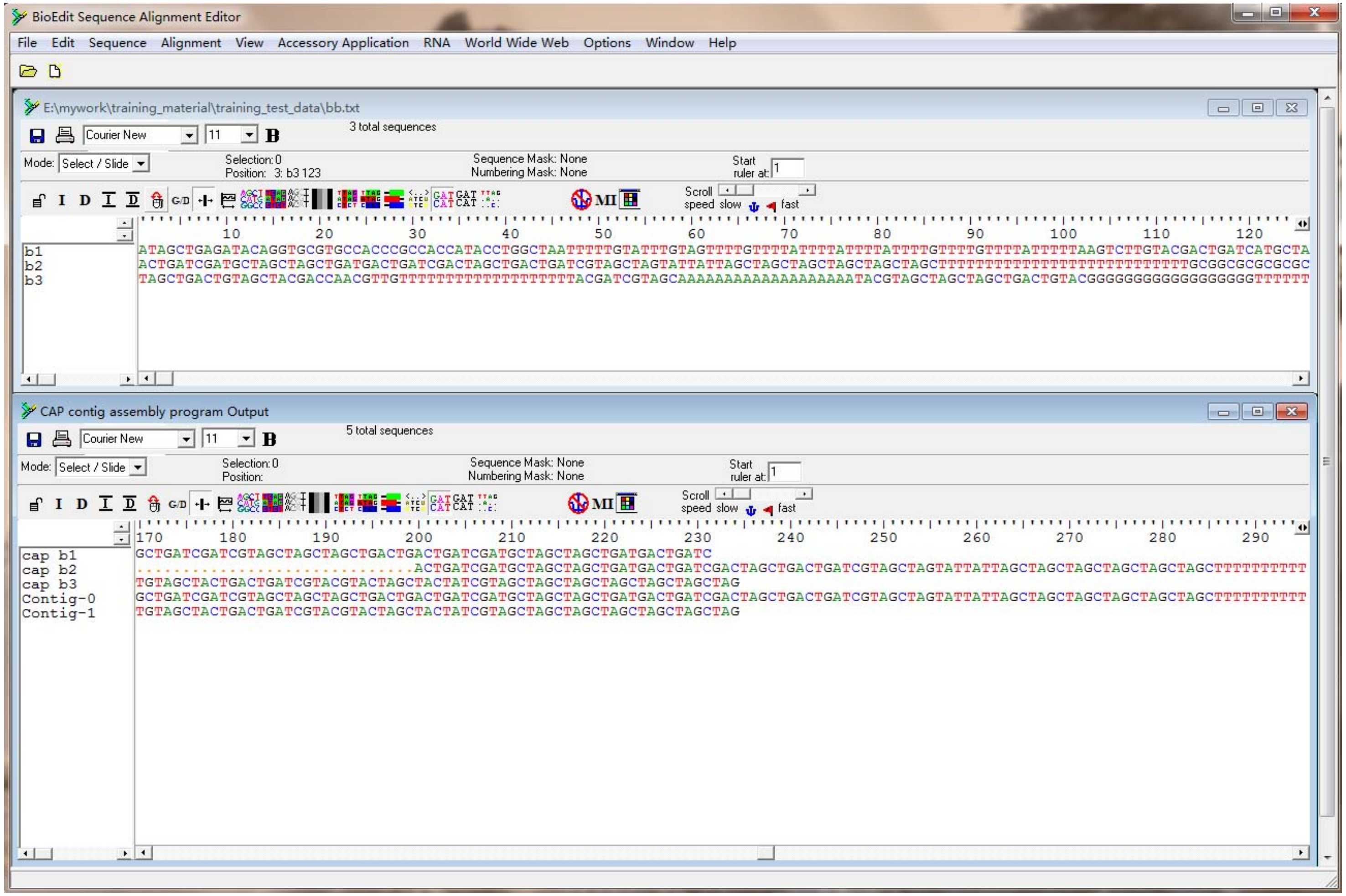
Map 4: Contig (HuRef) Table View
Region Displayed: **53,070K-53,077,800 bp** Download/View Sequence/Evidence
Total Contigs On Chromosome: **174**
Contigs Labeled: **1** Total Contigs in Region: **1**

Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

Sequence assembly

Software BioEdit

File -> open -> sequence -> Accessory Application -> contig assembly


BioEdit Sequence Alignment Editor
File Edit Sequence Alignment View Accessory Application RNA World Wide Web Options Window Help
E:\mywork\training_material\training_test_data\bb.txt
Courier New
11
B
Mode: Select / Slide
Scroll speed slow fast
CAP contig assembly program Output
Courier New
11
B
Mode: Select / Slide
Scroll speed slow fast

3 total sequences

Selection: 0
Position: 3: b3 123
Sequence Mask: None
Numbering Mask: None
Start ruler at: 1

```
         10        20        30        40        50        60        70        80        90        100       110       120
b1   ATAGCTGAGATACAGGTGCGTGCCACCCGCCACCATACCTGGCTAATTTTTGTATTTGTAGTTTTGTTTTATTTTATTTTATTTTGTTTTGTTTTATTTTTAAGTCTTGTACGACTGATCATGCTA
b2   ACTGATCGATGCTAGCTAGCTGATGACTGATCGACTAGCTGACTGATCGTAGCTAGTATTATTAGCTAGCTAGCTAGCTAGCTAGCTTTTTTTTTTTTTTTTTTTTTTTTTTTGCGGCGCGCGCGC
b3   TAGCTGACTGTAGCTACGACCAACGTTGTTTTTTTTTTTTTTTTTTACGATCGTAGCAAAAAAAAAAAAAAAAAAAATACGTAGCTAGCTAGCTGACTGTACGGGGGGGGGGGGGGGGGTTTTTT
```

5 total sequences

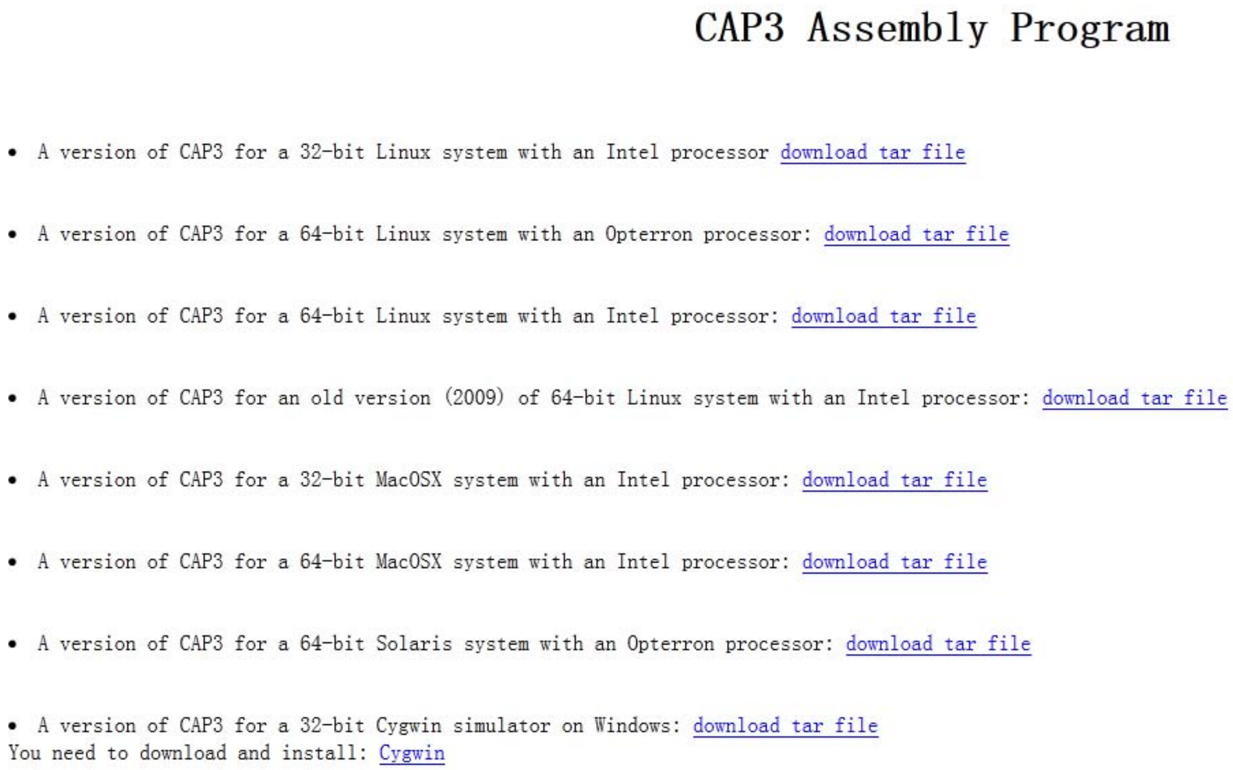
Selection: 0
Position:
Sequence Mask: None
Numbering Mask: None
Start ruler at: 1

```
          170       180       190       200       210       220       230       240       250       260       270       280       290
cap b1    GCTGATCGATCGTAGCTAGCTAGCTGACTGACTGATCGATGCTAGCTAGCTGATGACTGATC
cap b2    ..............................ACTGATCGATGCTAGCTAGCTGATGACTGATCGACTAGCTGACTGATCGTAGCTAGTATTATTAGCTAGCTAGCTAGCTAGCTAGCTTTTTTTTTT
cap b3    TGTAGCTACTGACTGATCGTACGTACTAGCTACTATCGTAGCTAGCTAGCTAGCTAGCTAG
Contig-0  GCTGATCGATCGTAGCTAGCTAGCTGACTGACTGATCGATGCTAGCTAGCTGATGACTGATCGACTAGCTGACTGATCGTAGCTAGTATTATTAGCTAGCTAGCTAGCTAGCTAGCTTTTTTTTTT
Contig-1  TGTAGCTACTGACTGATCGTACGTACTAGCTACTATCGTAGCTAGCTAGCTAGCTAGCTAG
```

Sequence assembly

CAP3 (contig assembly program)

http://seq.cs.iastate.edu/cap3.html



CAP3 Assembly Program

- A version of CAP3 for a 32-bit Linux system with an Intel processor download tar file
- A version of CAP3 for a 64-bit Linux system with an Opterron processor: download tar file
- A version of CAP3 for a 64-bit Linux system with an Intel processor: download tar file
- A version of CAP3 for an old version (2009) of 64-bit Linux system with an Intel processor: download tar file
- A version of CAP3 for a 32-bit MacOSX system with an Intel processor: download tar file
- A version of CAP3 for a 64-bit MacOSX system with an Intel processor: download tar file
- A version of CAP3 for a 64-bit Solaris system with an Opterron processor: download tar file
- A version of CAP3 for a 32-bit Cygwin simulator on Windows: download tar file
You need to download and install: Cygwin

PART Ⅲ insight into BLAST

Concept of BLAST

Concept: The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Purpose: predict function

infer evolutional tree

construct families

Concept of BLAST

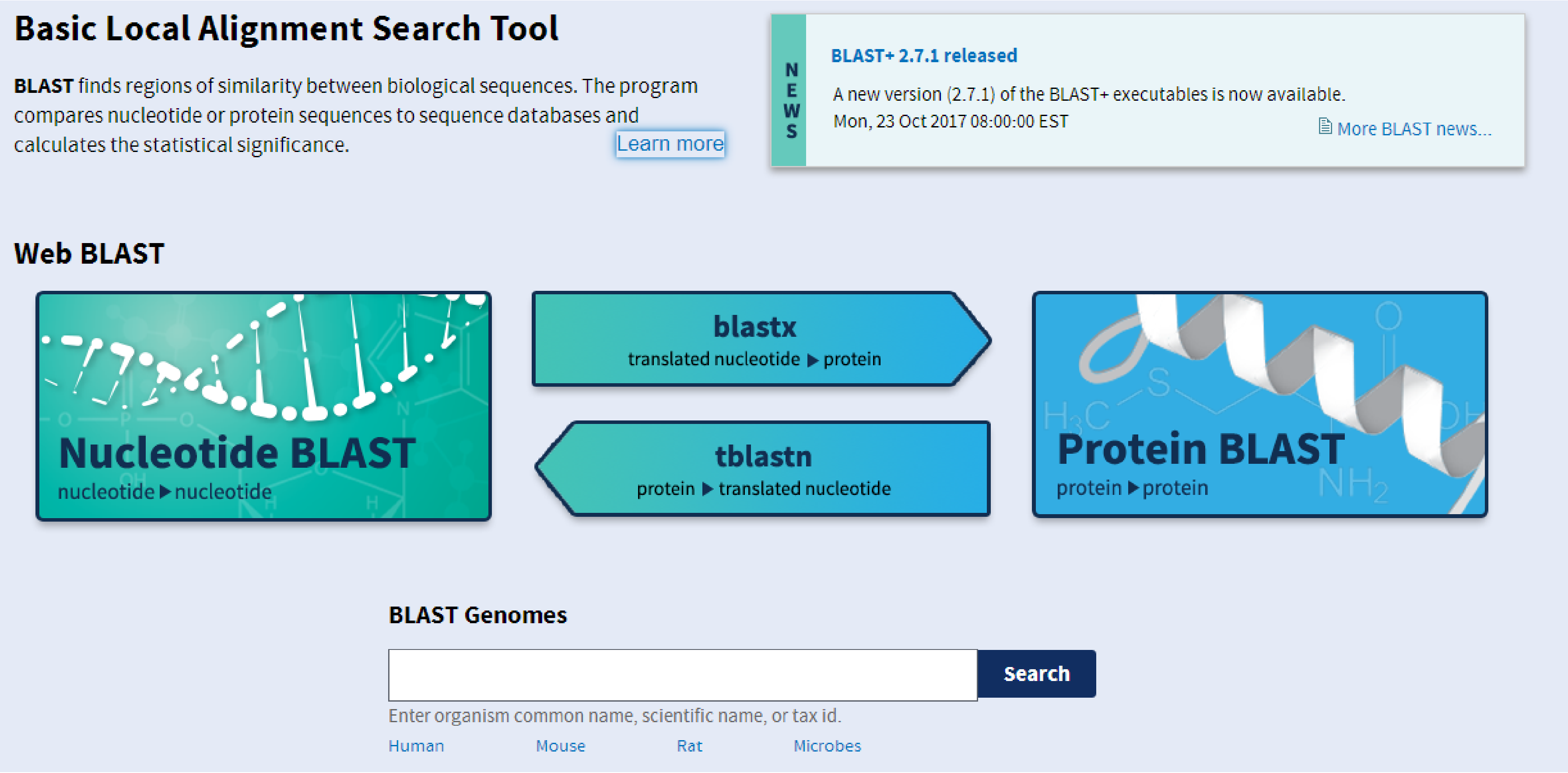
BLAST : word size -> **high-scoring segment (HSP)**

query sequence : (nucleotide/translated nucleotide, protein)

database: (nucleotide/translated nucleotide, protein)

BLAST HOME

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



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

BLAST+ 2.7.1 released

A new version (2.7.1) of the BLAST+ executables is now available.
Mon, 23 Oct 2017 08:00:00 EST

More BLAST news...

Web BLAST

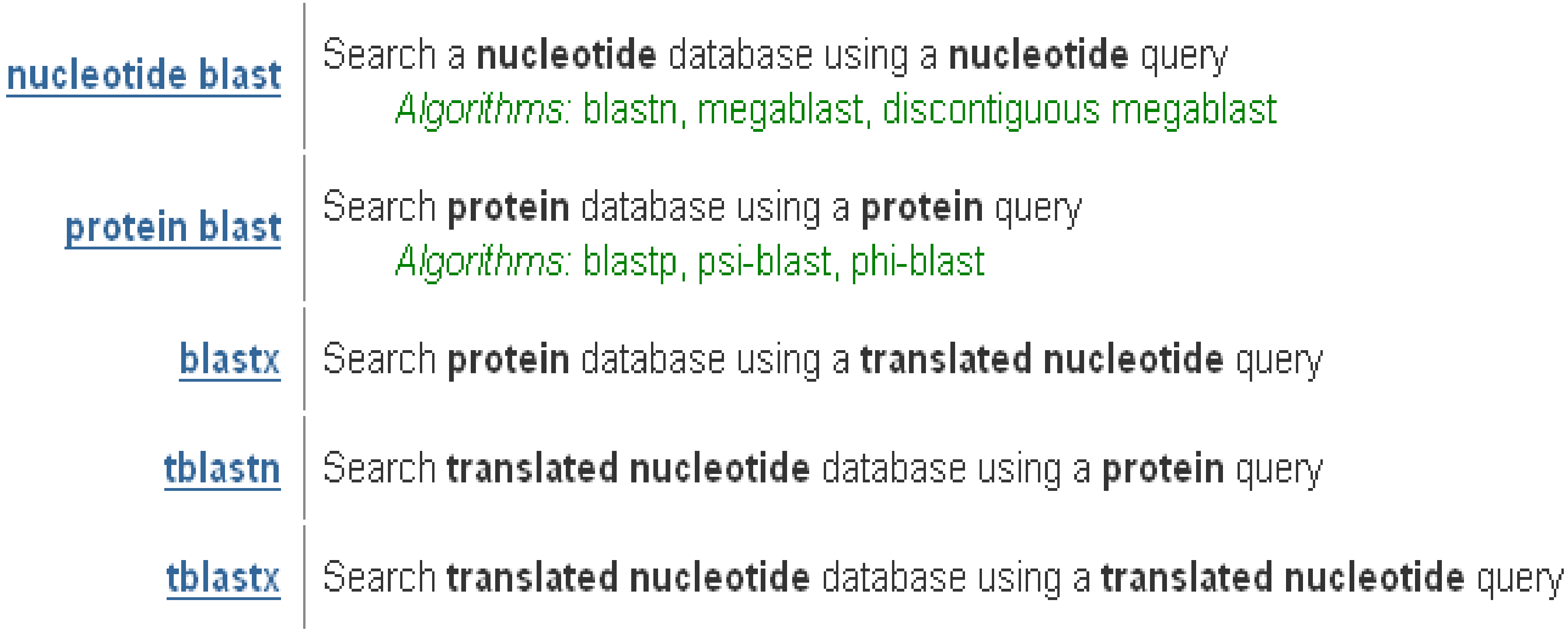
BLAST Genomes

Search

Enter organism common name, scientific name, or tax id.

Human Mouse Rat Microbes

Basic program of BLAST



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

PSI-BLAST

Position-Specific Iterated BLAST (PSI-BLAST, family protein)

(1) PSI-BLAST takes as an input a single protein sequence and compares it to a protein database, using the gapped BLAST program

(2) The program constructs a multiple alignment, and then a profile, from any significant local alignments found.

(3) The profile is compared to the protein database, again seeking local alignments.

(4) PSI-BLAST estimates the statistical significance of the local alignments found.

(5) Finally, PSI-BLAST iterates, by returning to step (2), an arbitrary number of times or until convergence.

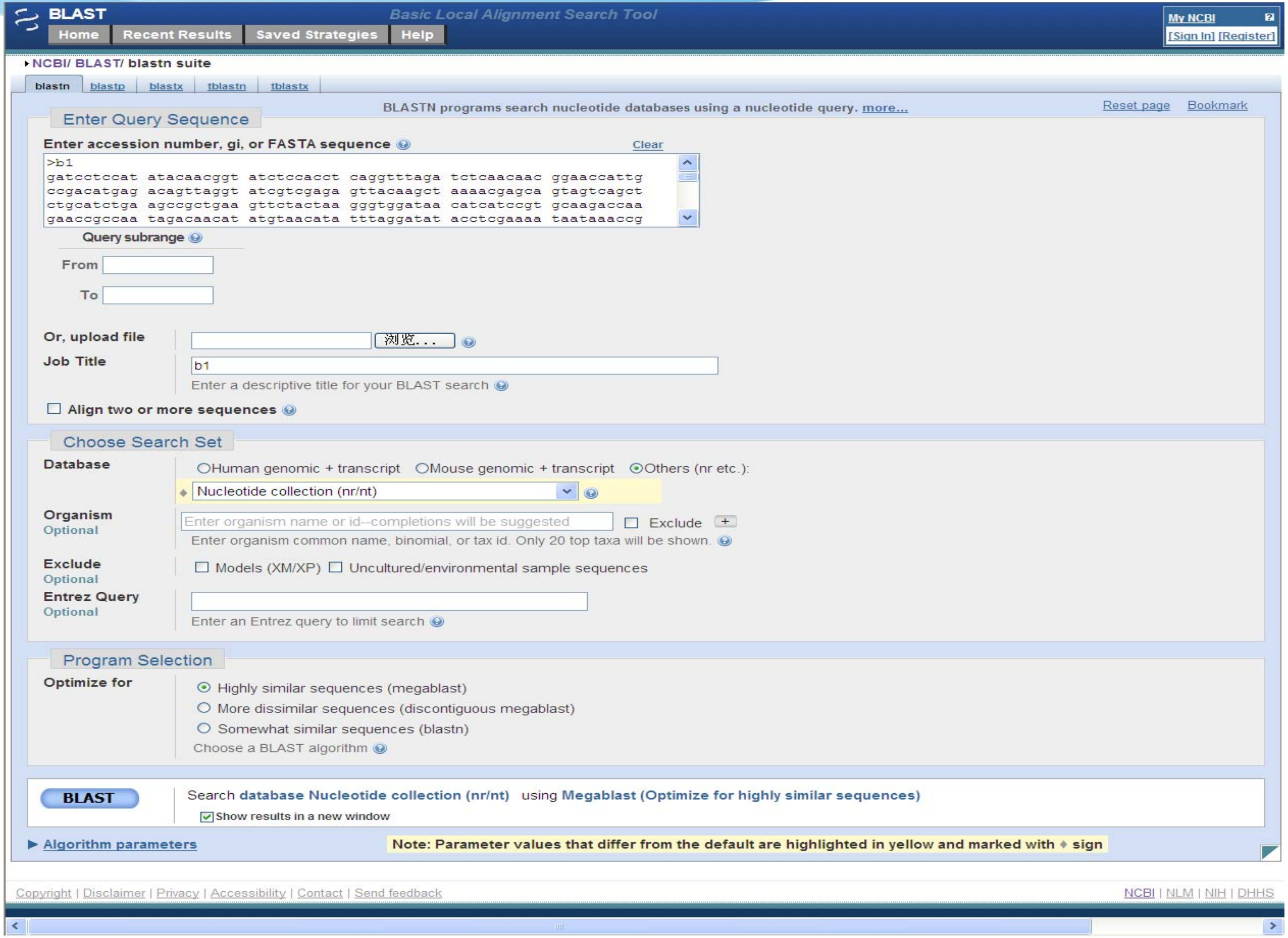
Summary of BLAST

- **Traditional BLAST (formerly blastall) nucleotide, protein, translations**
  - **blastn nucleotide query vs. nucleotide database**
  - **blastp protein query vs. protein database**
  - **blastx nucleotide query vs. protein database**
  - **tblastn protein query vs. translated nucleotide database**
  - **tblastx translated nucleotide query vs. translated nucleotide database**

- **Position Specific BLAST Programs protein only**
  - **Position Specific Iterative BLAST (PSI-BLAST)**

    Automatically generates a position specific score matrix (PSSM)

BLAST use – web means (nucleotide)



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

Enter Query Sequence

Enter accession number, gi, or FASTA sequence   Clear

```
>b1
gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
```

Query subrange

From

To

Or, upload file   浏览...

Job Title   b1

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

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

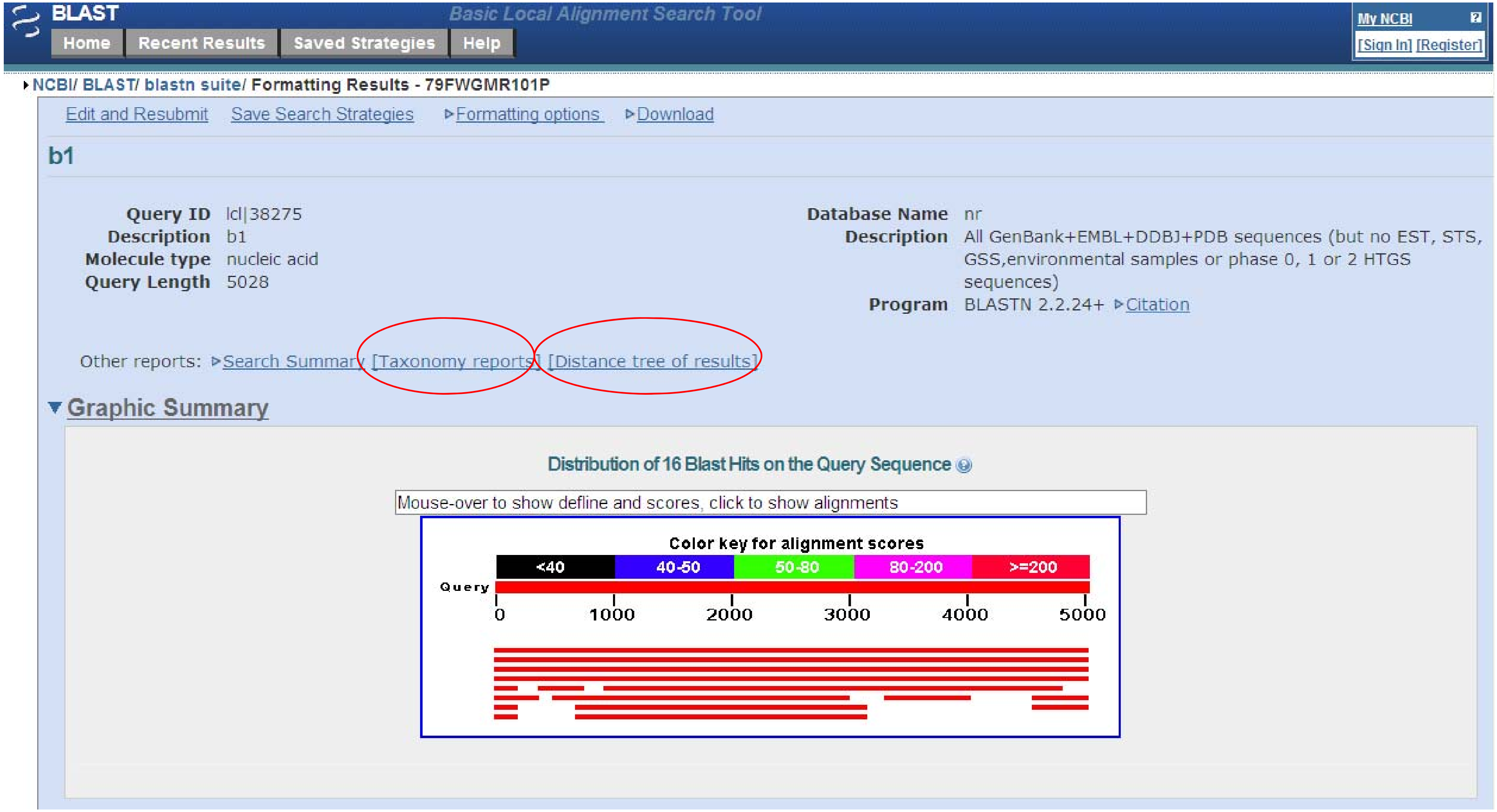
☑ Show results in a new window

Algorithm parameters

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

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback
NCBI | NLM | NIH | DHHS

BLAST use – web means (nucleotide)



NCBI/ BLAST/ blastn suite/ Formatting Results - 79FWGMR101P

Edit and Resubmit   Save Search Strategies   ▷Formatting options   ▷Download

b1

Query ID lcl|38275
Description b1
Molecule type nucleic acid
Query Length 5028

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 2.2.24+ ▷Citation

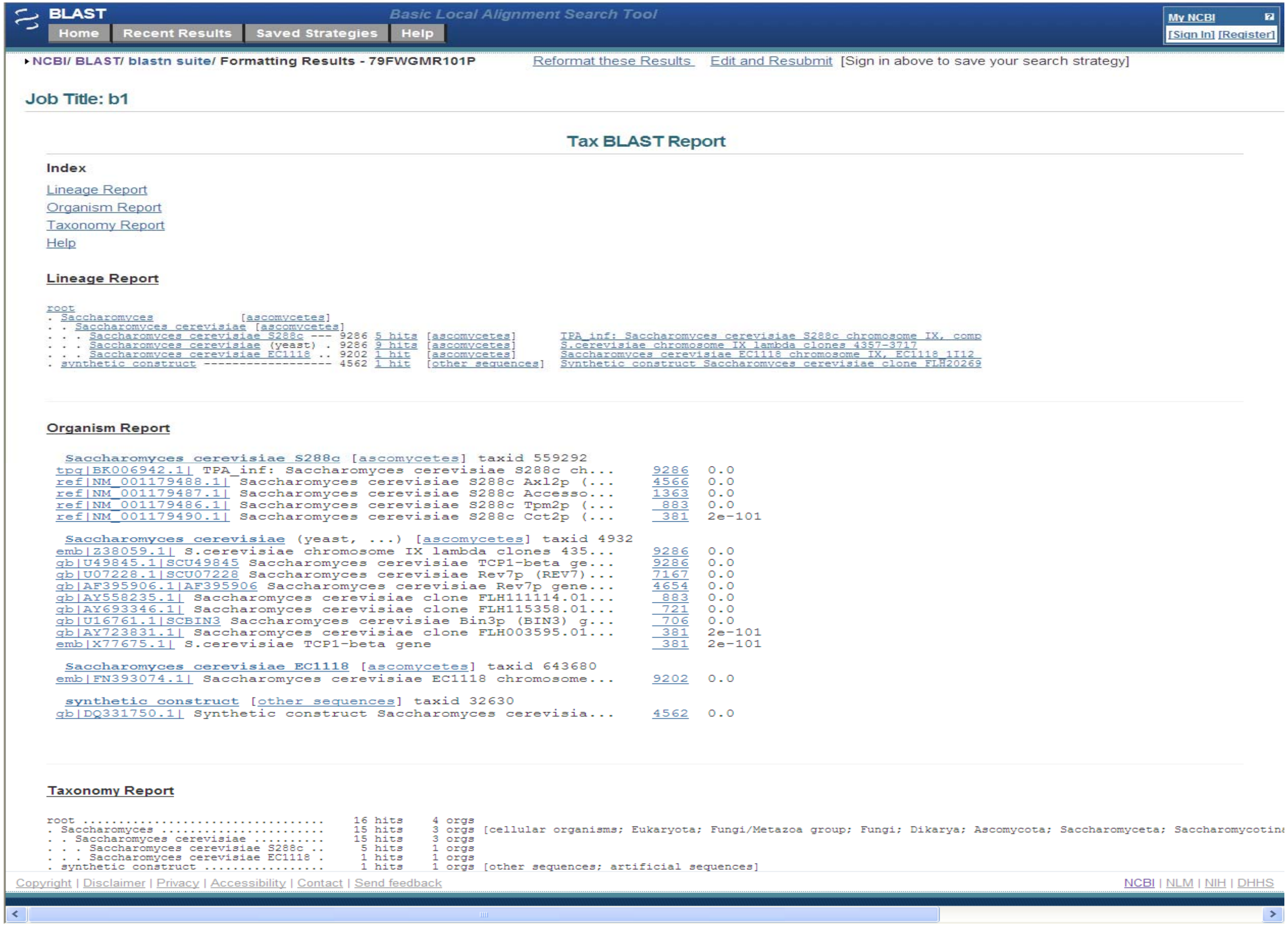
Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results]

▼ Graphic Summary

Distribution of 16 Blast Hits on the Query Sequence

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

BLAST use – web means (taxonomy report)



NCBI/ BLAST/ blastn suite/ Formatting Results - 79FWGMR101P   Reformat these Results   Edit and Resubmit [Sign in above to save your search strategy]

Job Title: b1

Tax BLAST Report

Index

Lineage Report
Organism Report
Taxonomy Report
Help

Lineage Report

```
root
. Saccharomyces              [ascomycetes]
. . Saccharomyces cerevisiae [ascomycetes]
. . . Saccharomyces cerevisiae S288c --- 9286 5 hits [ascomycetes]      TPA_inf: Saccharomyces cerevisiae S288c chromosome IX, comp
. . . Saccharomyces cerevisiae (yeast) . 9286 9 hits [ascomycetes]      S.cerevisiae chromosome IX lambda clones 4357-3717
. . . Saccharomyces cerevisiae EC1118 .. 9202 1 hit  [ascomycetes]      Saccharomyces cerevisiae EC1118 chromosome IX, EC1118_1I12
. synthetic construct ------------------ 4562 1 hit  [other sequences]  Synthetic construct Saccharomyces cerevisiae clone FLH20269
```

Organism Report

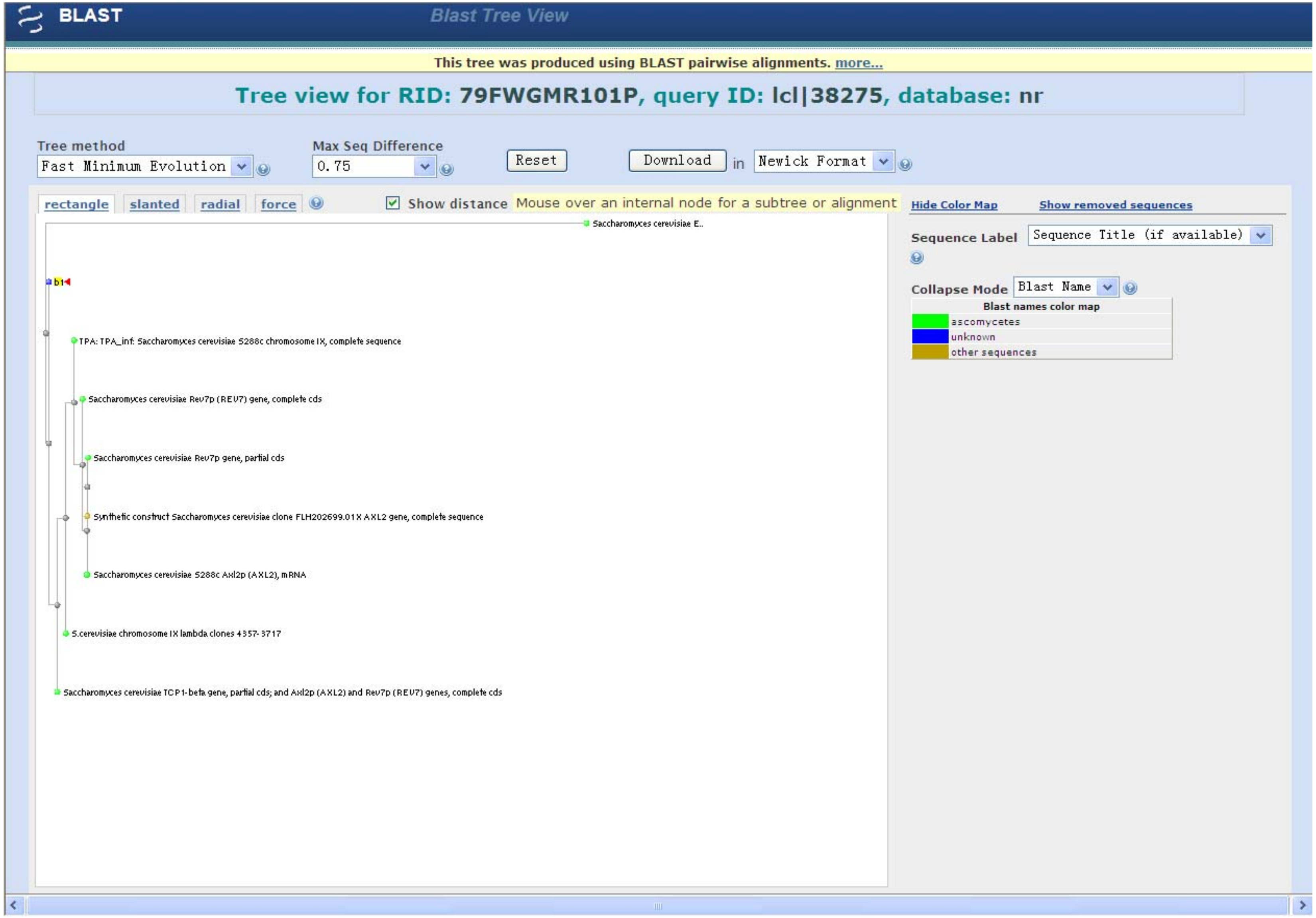
```
 Saccharomyces cerevisiae S288c [ascomycetes] taxid 559292
tpg|BK006942.1| TPA_inf: Saccharomyces cerevisiae S288c ch...     9286  0.0
ref|NM_001179488.1| Saccharomyces cerevisiae S288c Axl2p (...     4566  0.0
ref|NM_001179487.1| Saccharomyces cerevisiae S288c Accesso...     1363  0.0
ref|NM_001179486.1| Saccharomyces cerevisiae S288c Tpm2p (...      883  0.0
ref|NM_001179490.1| Saccharomyces cerevisiae S288c Cct2p (...      381  2e-101

 Saccharomyces cerevisiae (yeast, ...) [ascomycetes] taxid 4932
emb|Z38059.1| S.cerevisiae chromosome IX lambda clones 435...     9286  0.0
gb|U49845.1|SCU49845 Saccharomyces cerevisiae TCP1-beta ge...     9286  0.0
gb|U07228.1|SCU07228 Saccharomyces cerevisiae Rev7p (REV7)...     7167  0.0
gb|AF395906.1|AF395906 Saccharomyces cerevisiae Rev7p gene...     4654  0.0
gb|AY558235.1| Saccharomyces cerevisiae clone FLH111114.01...      883  0.0
gb|AY693346.1| Saccharomyces cerevisiae clone FLH115358.01...      721  0.0
gb|U16761.1|SCBIN3 Saccharomyces cerevisiae Bin3p (BIN3) g...      706  0.0
gb|AY723831.1| Saccharomyces cerevisiae clone FLH003595.01...      381  2e-101
emb|X77675.1| S.cerevisiae TCP1-beta gene                          381  2e-101

 Saccharomyces cerevisiae EC1118 [ascomycetes] taxid 643680
emb|FN393074.1| Saccharomyces cerevisiae EC1118 chromosome...     9202  0.0

 synthetic construct [other sequences] taxid 32630
gb|DQ331750.1| Synthetic construct Saccharomyces cerevisia...     4562  0.0
```

Taxonomy Report

```
root ..................................   16 hits    4 orgs
. Saccharomyces .......................   15 hits    3 orgs [cellular organisms; Eukaryota; Fungi/Metazoa group; Fungi; Dikarya; Ascomycota; Saccharomyceta; Saccharomycotina
. . Saccharomyces cerevisiae ..........   15 hits    3 orgs
. . . Saccharomyces cerevisiae S288c ..    5 hits    1 orgs
. . . Saccharomyces cerevisiae EC1118 .    1 hits    1 orgs
. synthetic construct .................    1 hits    1 orgs [other sequences; artificial sequences]
```

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback
NCBI | NLM | NIH | DHHS

BLAST use – infer evolution tree



BLAST Blast Tree View

This tree was produced using BLAST pairwise alignments. more...

Tree view for RID: 79FWGMR101P, query ID: lcl|38275, database: nr

Tree method: Fast Minimum Evolution

Max Seq Difference: 0.75

Download in Newick Format

☑ Show distance

Mouse over an internal node for a subtree or alignment

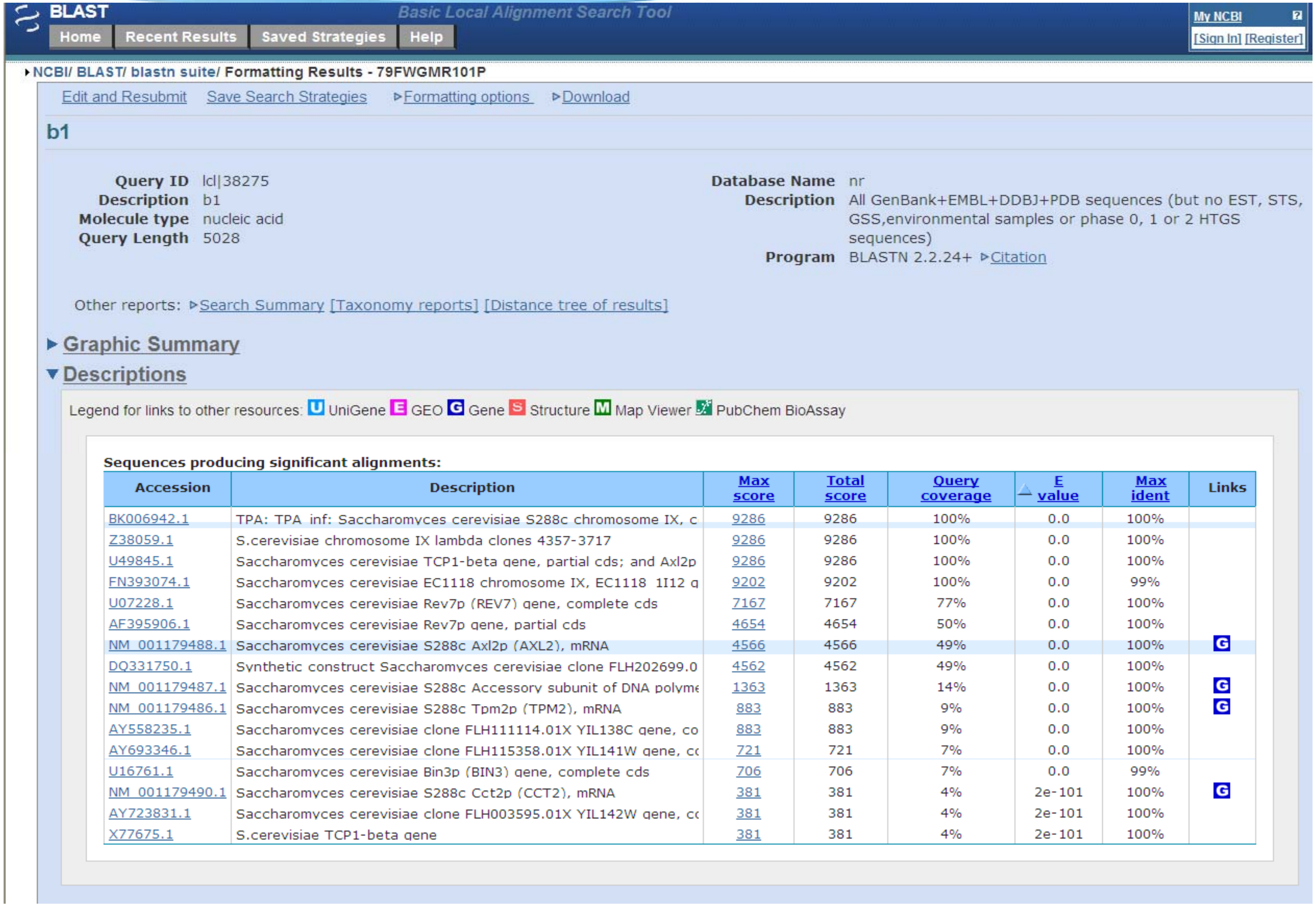
Sequence Label: Sequence Title (if available)

Collapse Mode: Blast Name

Blast names color map	
🟩	ascomycetes
🟦	unknown
🟨	other sequences

BLAST
Reset
Download
rectangle
slanted
radial
force
Hide Color Map
Show removed sequences
Saccharomyces cerevisiae E..
b1
TPA: TPA_inf: Saccharomyces cerevisiae S288c chromosome IX, complete sequence
Saccharomyces cerevisiae Rev7p (REV7) gene, complete cds
Saccharomyces cerevisiae Rev7p gene, partial cds
Synthetic construct Saccharomyces cerevisiae clone FLH202699.01X AXL2 gene, complete sequence
Saccharomyces cerevisiae S288c Axl2p (AXL2), mRNA
S.cerevisiae chromosome IX lambda clones 4357-3717
Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p (AXL2) and Rev7p (REV7) genes, complete cds

BLAST use – predict function



NCBI/ BLAST/ blastn suite/ Formatting Results - 79FWGMR101P

Edit and Resubmit   Save Search Strategies   ▹Formatting options   ▹Download

b1

Query ID lcl|38275
Description b1
Molecule type nucleic acid
Query Length 5028

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 2.2.24+ ▹Citation

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

▸ Graphic Summary

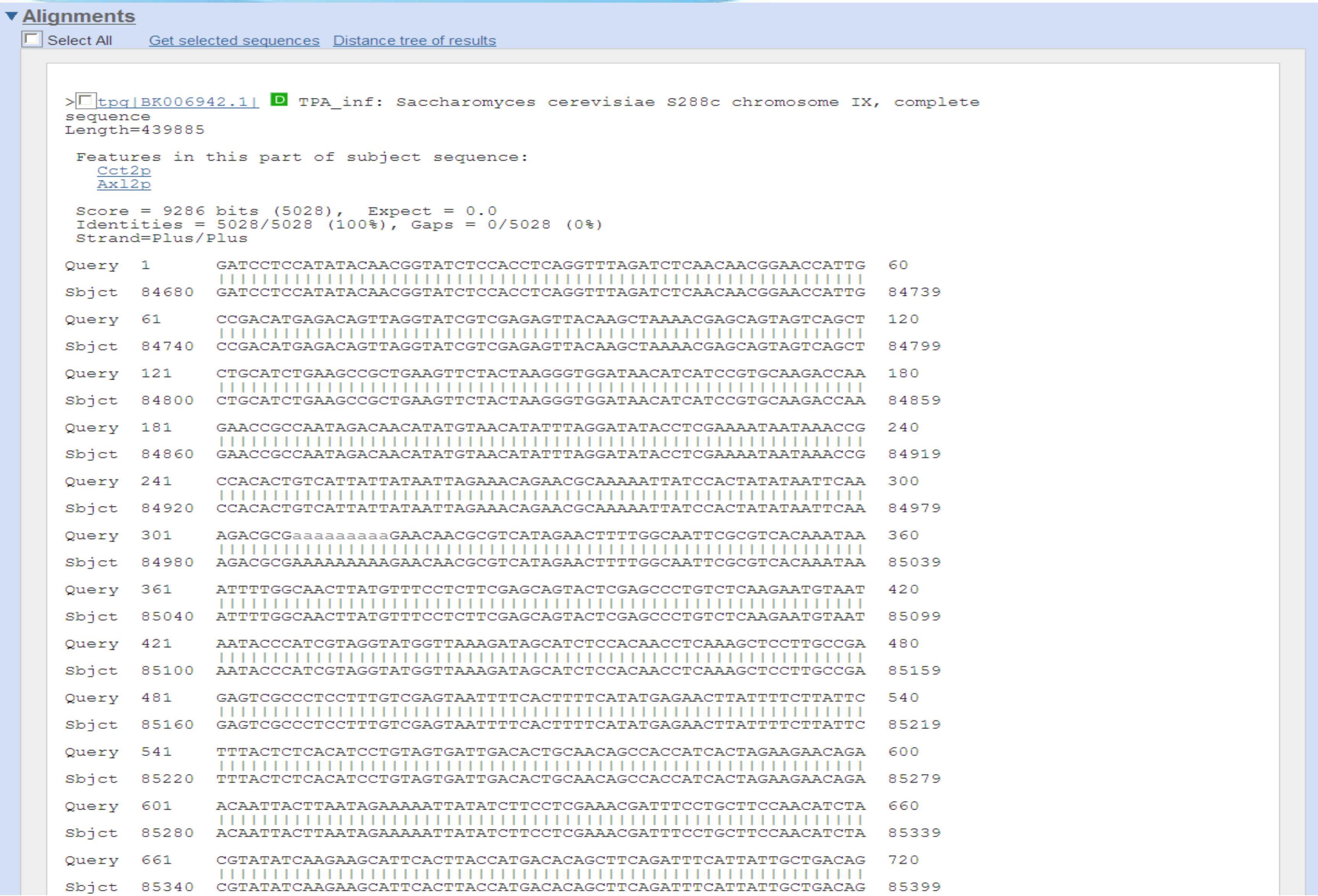
▾ Descriptions

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

Sequences producing significant alignments:

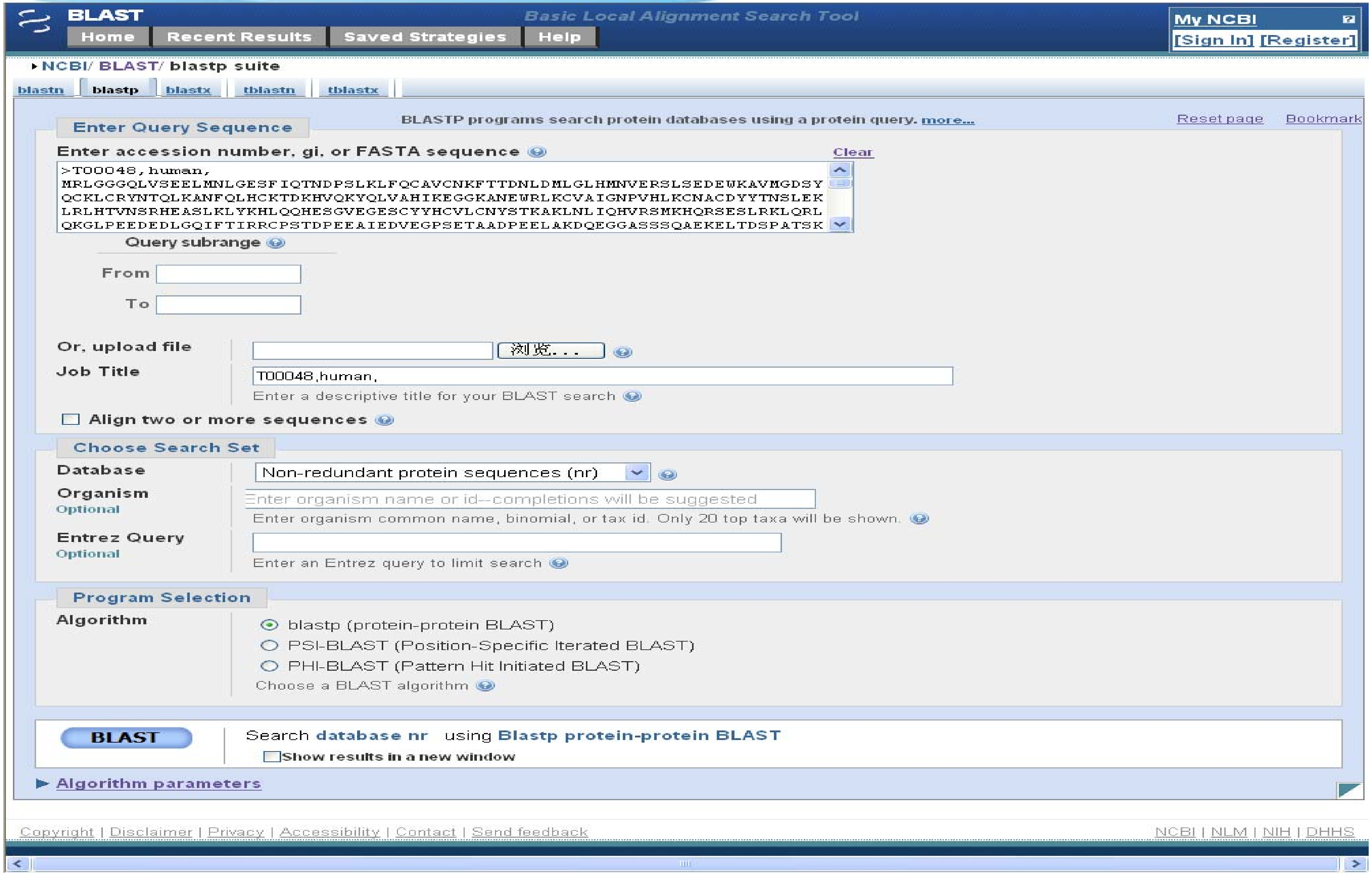
Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
BK006942.1	TPA: TPA inf: Saccharomyces cerevisiae S288c chromosome IX, c	9286	9286	100%	0.0	100%	
Z38059.1	S.cerevisiae chromosome IX lambda clones 4357-3717	9286	9286	100%	0.0	100%	
U49845.1	Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p	9286	9286	100%	0.0	100%	
FN393074.1	Saccharomyces cerevisiae EC1118 chromosome IX, EC1118_1I12_g	9202	9202	100%	0.0	99%	
U07228.1	Saccharomyces cerevisiae Rev7p (REV7) gene, complete cds	7167	7167	77%	0.0	100%	
AF395906.1	Saccharomyces cerevisiae Rev7p gene, partial cds	4654	4654	50%	0.0	100%	
NM_001179488.1	Saccharomyces cerevisiae S288c Axl2p (AXL2), mRNA	4566	4566	49%	0.0	100%	G
DQ331750.1	Synthetic construct Saccharomyces cerevisiae clone FLH202699.0	4562	4562	49%	0.0	100%	
NM_001179487.1	Saccharomyces cerevisiae S288c Accessory subunit of DNA polyme	1363	1363	14%	0.0	100%	G
NM_001179486.1	Saccharomyces cerevisiae S288c Tpm2p (TPM2), mRNA	883	883	9%	0.0	100%	G
AY558235.1	Saccharomyces cerevisiae clone FLH111114.01X YIL138C gene, co	883	883	9%	0.0	100%	
AY693346.1	Saccharomyces cerevisiae clone FLH115358.01X YIL141W gene, cc	721	721	7%	0.0	100%	
U16761.1	Saccharomyces cerevisiae Bin3p (BIN3) gene, complete cds	706	706	7%	0.0	99%	
NM_001179490.1	Saccharomyces cerevisiae S288c Cct2p (CCT2), mRNA	381	381	4%	2e-101	100%	G
AY723831.1	Saccharomyces cerevisiae clone FLH003595.01X YIL142W gene, cc	381	381	4%	2e-101	100%	
X77675.1	S.cerevisiae TCP1-beta gene	381	381	4%	2e-101	100%	

BLAST use – web manners (alignment report)



```
>□tpg|BK006942.1| D TPA_inf: Saccharomyces cerevisiae S288c chromosome IX, complete
sequence
Length=439885

 Features in this part of subject sequence:
   Cct2p
   Axl2p

 Score = 9286 bits (5028),  Expect = 0.0
 Identities = 5028/5028 (100%), Gaps = 0/5028 (0%)
 Strand=Plus/Plus

Query  1      GATCCTCCATATACAACGGTATCTCCACCTCAGGTTTAGATCTCAACAACGGAACCATTG  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  84680  GATCCTCCATATACAACGGTATCTCCACCTCAGGTTTAGATCTCAACAACGGAACCATTG  84739

Query  61     CCGACATGAGACAGTTAGGTATCGTCGAGAGTTACAAGCTAAAACGAGCAGTAGTCAGCT  120
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  84740  CCGACATGAGACAGTTAGGTATCGTCGAGAGTTACAAGCTAAAACGAGCAGTAGTCAGCT  84799

Query  121    CTGCATCTGAAGCCGCTGAAGTTCTACTAAGGGTGGATAACATCATCCGTGCAAGACCAA  180
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  84800  CTGCATCTGAAGCCGCTGAAGTTCTACTAAGGGTGGATAACATCATCCGTGCAAGACCAA  84859

Query  181    GAACCGCCAATAGACAACATATGTAACATATTTAGGATATACCTCGAAAATAATAAACCG  240
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  84860  GAACCGCCAATAGACAACATATGTAACATATTTAGGATATACCTCGAAAATAATAAACCG  84919

Query  241    CCACACTGTCATTATTATAATTAGAAACAGAACGCAAAAATTATCCACTATATAATTCAA  300
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  84920  CCACACTGTCATTATTATAATTAGAAACAGAACGCAAAAATTATCCACTATATAATTCAA  84979

Query  301    AGACGCGaaaaaaaaaGAACAACGCGTCATAGAACTTTTGGCAATTCGCGTCACAAATAA  360
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  84980  AGACGCGAAAAAAAAAGAACAACGCGTCATAGAACTTTTGGCAATTCGCGTCACAAATAA  85039

Query  361    ATTTTGGCAACTTATGTTTCCTCTTCGAGCAGTACTCGAGCCCTGTCTCAAGAATGTAAT  420
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  85040  ATTTTGGCAACTTATGTTTCCTCTTCGAGCAGTACTCGAGCCCTGTCTCAAGAATGTAAT  85099

Query  421    AATACCCATCGTAGGTATGGTTAAAGATAGCATCTCCACAACCTCAAAGCTCCTTGCCGA  480
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  85100  AATACCCATCGTAGGTATGGTTAAAGATAGCATCTCCACAACCTCAAAGCTCCTTGCCGA  85159

Query  481    GAGTCGCCCTCCTTTGTCGAGTAATTTTCACTTTTCATATGAGAACTTATTTTCTTATTC  540
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  85160  GAGTCGCCCTCCTTTGTCGAGTAATTTTCACTTTTCATATGAGAACTTATTTTCTTATTC  85219

Query  541    TTTACTCTCACATCCTGTAGTGATTGACACTGCAACAGCCACCATCACTAGAAGAACAGA  600
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  85220  TTTACTCTCACATCCTGTAGTGATTGACACTGCAACAGCCACCATCACTAGAAGAACAGA  85279

Query  601    ACAATTACTTAATAGAAAAATTATATCTTCCTCGAAACGATTTCCTGCTTCCAACATCTA  660
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  85280  ACAATTACTTAATAGAAAAATTATATCTTCCTCGAAACGATTTCCTGCTTCCAACATCTA  85339

Query  661    CGTATATCAAGAAGCATTCACTTACCATGACACAGCTTCAGATTTCATTATTGCTGACAG  720
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  85340  CGTATATCAAGAAGCATTCACTTACCATGACACAGCTTCAGATTTCATTATTGCTGACAG  85399
```

BLAST use – web manners (protein)



NCBI/ BLAST/ blastp suite

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

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

```
>T00048,human,
MRLGGGQLVSEELMNLGESFIQTNDPSLKLFQCAVCNKFTTDNLDMLGLHMNVERSLSEDEWKAVMGDSY
QCKLCRYNTQLKANFQLHCKTDKHVQKYQLVAHIKEGGKANEWRLKCVAIGNPVHLKCNACDYYTNSLEK
LRLHTVNSRHEASLKLYKHLQQHESGVEGESCYYHCVLCNYSTKAKLNLIQHVRSMKHQRSESLRKLQRL
QKGLPEEDEDLGQIFTIRRCPSTDPEEAIEDVEGPSETAADPEELAKDQEGGASSSQAEKELTDSPATSK
```

Query subrange

From

To

Or, upload file

Job Title: T00048,human,

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

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

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)

Choose a BLAST algorithm

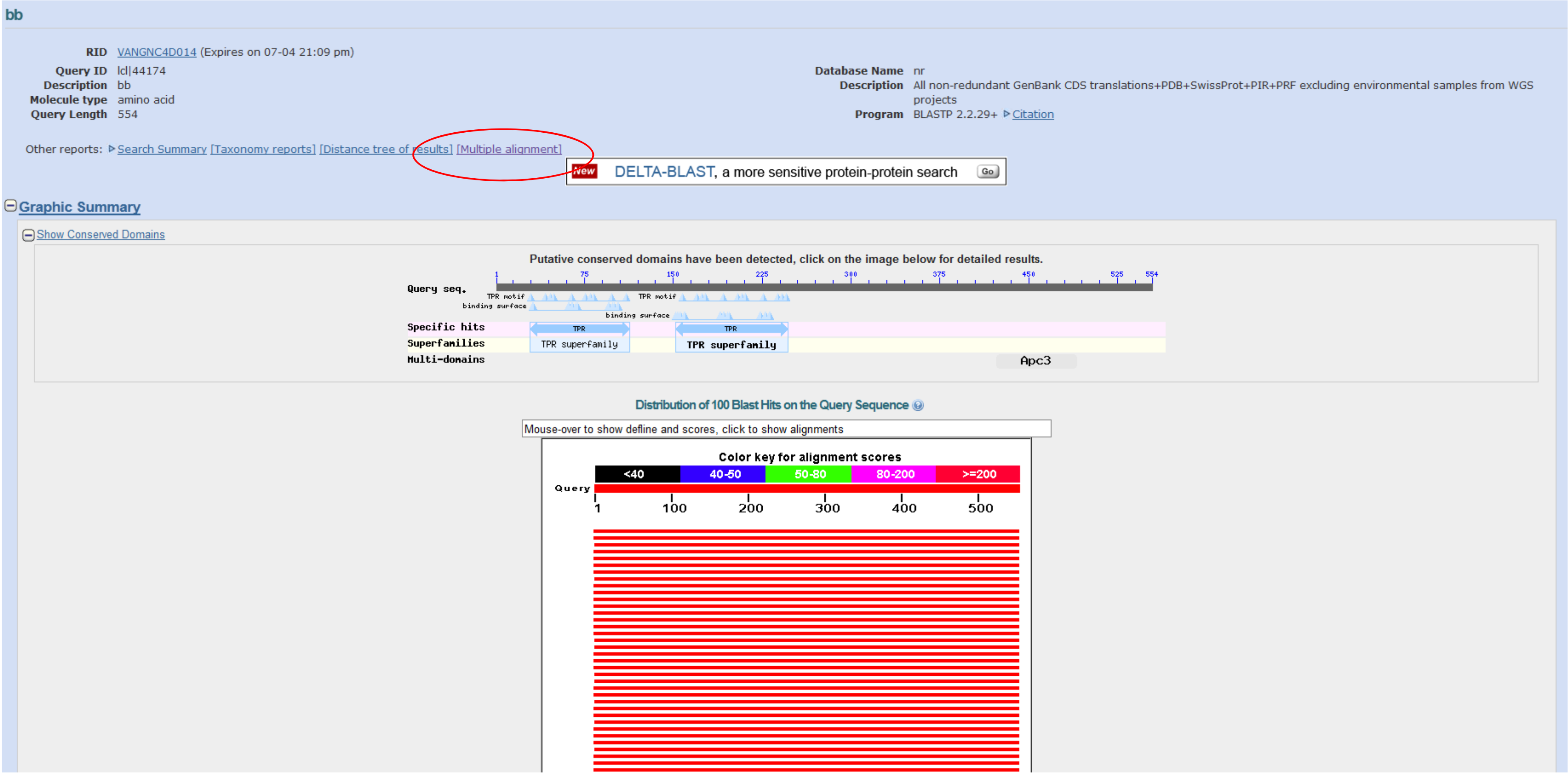
Search database nr using Blastp protein-protein BLAST

□ Show results in a new window

▶ Algorithm parameters

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback
NCBI | NLM | NIH | DHHS

BLAST use – result summary



bb

RID VANGNC4D014 (Expires on 07-04 21:09 pm)

Query ID lcl|44174
Description bb
Molecule type amino acid
Query Length 554

Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.2.29+ ▷ Citation

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

New DELTA-BLAST, a more sensitive protein-protein search Go

⊟ Graphic Summary

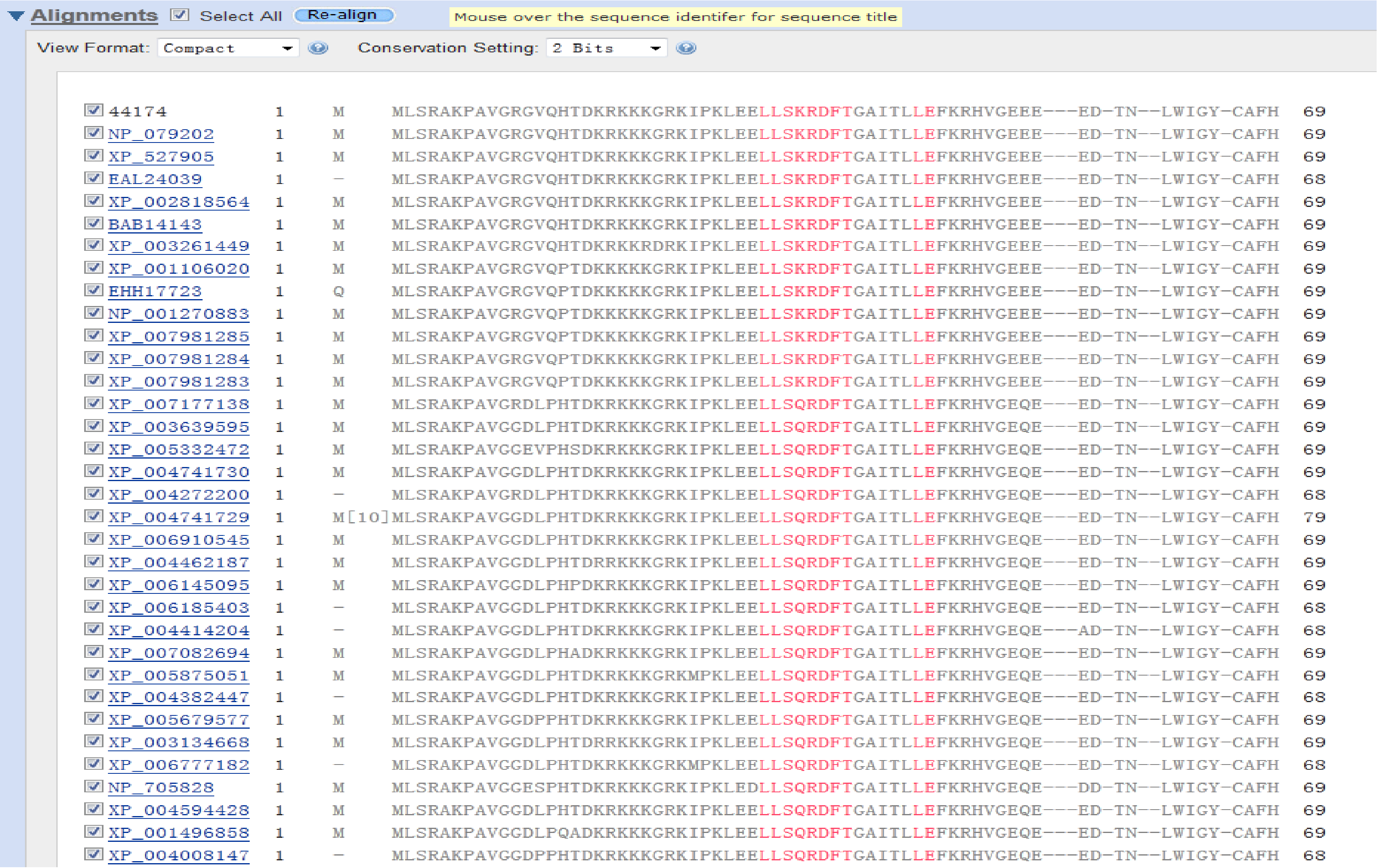
⊟ Show Conserved Domains

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

Distribution of 100 Blast Hits on the Query Sequence

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

BLAST use – multi alignment

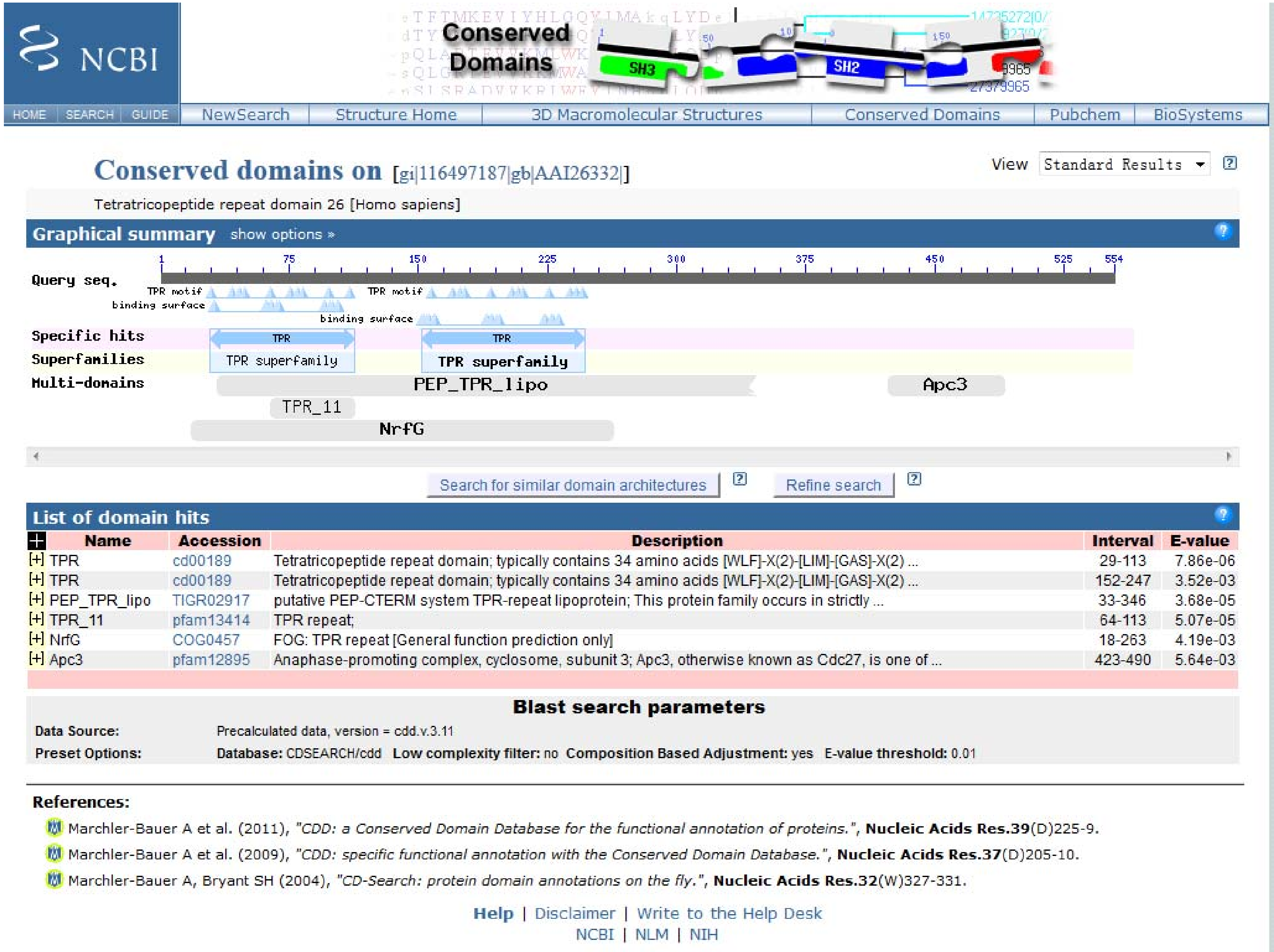


▼ Alignments ☑ Select All Re-align
Mouse over the sequence identifer for sequence title

View Format: Compact   Conservation Setting: 2 Bits

```
☑ 44174          1  M      MLSRAKPAVGRGVQHTDKRKKKGRKIPKLEELLSKRDFTGAITLLEFKRHVGEEE---ED-TN--LWIGY-CAFH  69
☑ NP_079202      1  M      MLSRAKPAVGRGVQHTDKRKKKGRKIPKLEELLSKRDFTGAITLLEFKRHVGEEE---ED-TN--LWIGY-CAFH  69
☑ XP_527905      1  M      MLSRAKPAVGRGVQHTDKRKKKGRKIPKLEELLSKRDFTGAITLLEFKRHVGEEE---ED-TN--LWIGY-CAFH  69
☑ EAL24039       1  -      MLSRAKPAVGRGVQHTDKRKKKGRKIPKLEELLSKRDFTGAITLLEFKRHVGEEE---ED-TN--LWIGY-CAFH  68
☑ XP_002818564   1  M      MLSRAKPAVGRGVQHTDKRKKKGRKIPKLEELLSKRDFTGAITLLEFKRHVGEEE---ED-TN--LWIGY-CAFH  69
☑ BAB14143       1  M      MLSRAKPAVGRGVQHTDKRKKKGRKIPKLEELLSKRDFTGAITLLEFKRHVGEEE---ED-TN--LWIGY-CAFH  69
☑ XP_003261449   1  M      MLSRAKPAVGRGVQHTDKRKKRDRKIPKLEELLSKRDFTGAITLLEFKRHVGEEE---ED-TN--LWIGY-CAFH  69
☑ XP_001106020   1  M      MLSRAKPAVGRGVQPTDKKKKKGRKIPKLEELLSKRDFTGAITLLEFKRHVGEEE---ED-TN--LWIGY-CAFH  69
☑ EHH17723       1  Q      MLSRAKPAVGRGVQPTDKKKKKGRKIPKLEELLSKRDFTGAITLLEFKRHVGEEE---ED-TN--LWIGY-CAFH  69
☑ NP_001270883   1  M      MLSRAKPAVGRGVQPTDKKKKKGRKIPKLEELLSKRDFTGAITLLEFKRHVGEEE---ED-TN--LWIGY-CAFH  69
☑ XP_007981285   1  M      MLSRAKPAVGRGVQPTDKKKKKGRKIPKLEELLSKRDFTGAITLLEFKRHVGEEE---ED-TN--LWIGY-CAFH  69
☑ XP_007981284   1  M      MLSRAKPAVGRGVQPTDKKKKKGRKIPKLEELLSKRDFTGAITLLEFKRHVGEEE---ED-TN--LWIGY-CAFH  69
☑ XP_007981283   1  M      MLSRAKPAVGRGVQPTDKKKKKGRKIPKLEELLSKRDFTGAITLLEFKRHVGEEE---ED-TN--LWIGY-CAFH  69
☑ XP_007177138   1  M      MLSRAKPAVGRDLPHTDKRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  69
☑ XP_003639595   1  M      MLSRAKPAVGGDLPHTDKRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  69
☑ XP_005332472   1  M      MLSRAKPAVGGEVPHSDKRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  69
☑ XP_004741730   1  M      MLSRAKPAVGGDLPHTDKRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  69
☑ XP_004272200   1  -      MLSRAKPAVGRDLPHTDKRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  68
☑ XP_004741729   1  M[10]MLSRAKPAVGGDLPHTDKRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  79
☑ XP_006910545   1  M      MLSRAKPAVGGDLPHTDKRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  69
☑ XP_004462187   1  M      MLSRAKPAVGGDLPHTDRRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  69
☑ XP_006145095   1  M      MLSRAKPAVGGDLPHPDKRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  69
☑ XP_006185403   1  -      MLSRAKPAVGGDLPHTDKRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  68
☑ XP_004414204   1  -      MLSRAKPAVGGDLPHTDKRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---AD-TN--LWIGY-CAFH  68
☑ XP_007082694   1  M      MLSRAKPAVGGDLPHADKRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  69
☑ XP_005875051   1  M      MLSRAKPAVGGDLPHTDKRKKKGRKMPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  69
☑ XP_004382447   1  -      MLSRAKPAVGGDLPHTDKRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  68
☑ XP_005679577   1  M      MLSRAKPAVGGDPPHTDKRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  69
☑ XP_003134668   1  M      MLSRAKPAVGGDLPHTDRRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  69
☑ XP_006777182   1  -      MLSRAKPAVGGDLPHTDKRKKKGRKMPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  68
☑ NP_705828      1  M      MLSRAKPAVGGESPHTDKRKKKGRKIPKLEDLLSQRDFTGAITLLEFKRHVGEQE---DD-TN--LWIGY-CAFH  69
☑ XP_004594428   1  M      MLSRAKPAVGGDLPHTDKRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  69
☑ XP_001496858   1  M      MLSRAKPAVGGDLPQADKRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  69
☑ XP_004008147   1  -      MLSRAKPAVGGDPPHTDKRKKKGRKIPKLEELLSQRDFTGAITLLEFKRHVGEQE---ED-TN--LWIGY-CAFH  68
```

BLAST use – conserved domain



Conserved domains on [gi|116497187|gb|AAI26332|]

View Standard Results

Tetratricopeptide repeat domain 26 [Homo sapiens]

Graphical summary show options »

List of domain hits

	Name	Accession	Description	Interval	E-value
[+]	TPR	cd00189	Tetratricopeptide repeat domain; typically contains 34 amino acids [WLF]-X(2)-[LIM]-[GAS]-X(2) ...	29-113	7.86e-06
[+]	TPR	cd00189	Tetratricopeptide repeat domain; typically contains 34 amino acids [WLF]-X(2)-[LIM]-[GAS]-X(2) ...	152-247	3.52e-03
[+]	PEP_TPR_lipo	TIGR02917	putative PEP-CTERM system TPR-repeat lipoprotein; This protein family occurs in strictly ...	33-346	3.68e-05
[+]	TPR_11	pfam13414	TPR repeat;	64-113	5.07e-05
[+]	NrfG	COG0457	FOG: TPR repeat [General function prediction only]	18-263	4.19e-03
[+]	Apc3	pfam12895	Anaphase-promoting complex, cyclosome, subunit 3; Apc3, otherwise known as Cdc27, is one of ...	423-490	5.64e-03

Blast search parameters

Data Source: Precalculated data, version = cdd.v.3.11

Preset Options: **Database:** CDSEARCH/cdd **Low complexity filter:** no **Composition Based Adjustment:** yes **E-value threshold:** 0.01

References:

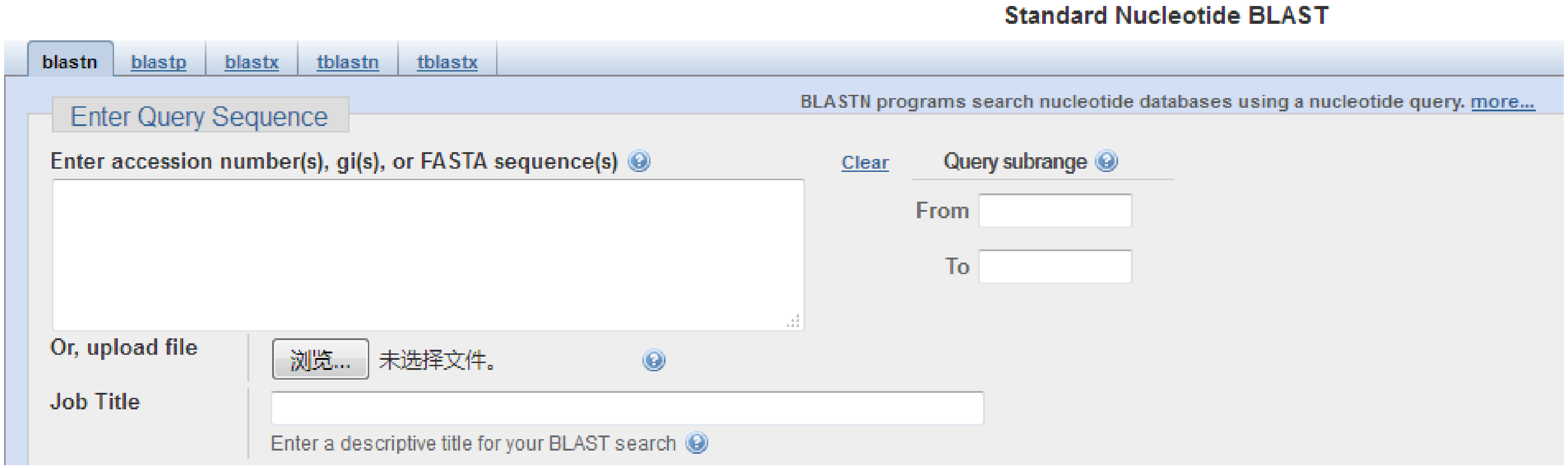
- Marchler-Bauer A et al. (2011), *"CDD: a Conserved Domain Database for the functional annotation of proteins."*, **Nucleic Acids Res.39**(D)225-9.
- Marchler-Bauer A et al. (2009), *"CDD: specific functional annotation with the Conserved Domain Database."*, **Nucleic Acids Res.37**(D)205-10.
- Marchler-Bauer A, Bryant SH (2004), *"CD-Search: protein domain annotations on the fly."*, **Nucleic Acids Res.32**(W)327-331.

Help | Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

BLAST use

Batch BLAST jobs

(1) input "batches" of sequences into one form and retrieve the results



Standard Nucleotide BLAST

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

Enter Query Sequence

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

Query subrange

From

To

Or, upload file

未选择文件。

Job Title

Enter a descriptive title for your BLAST search

BLAST use

Batch BLAST jobs

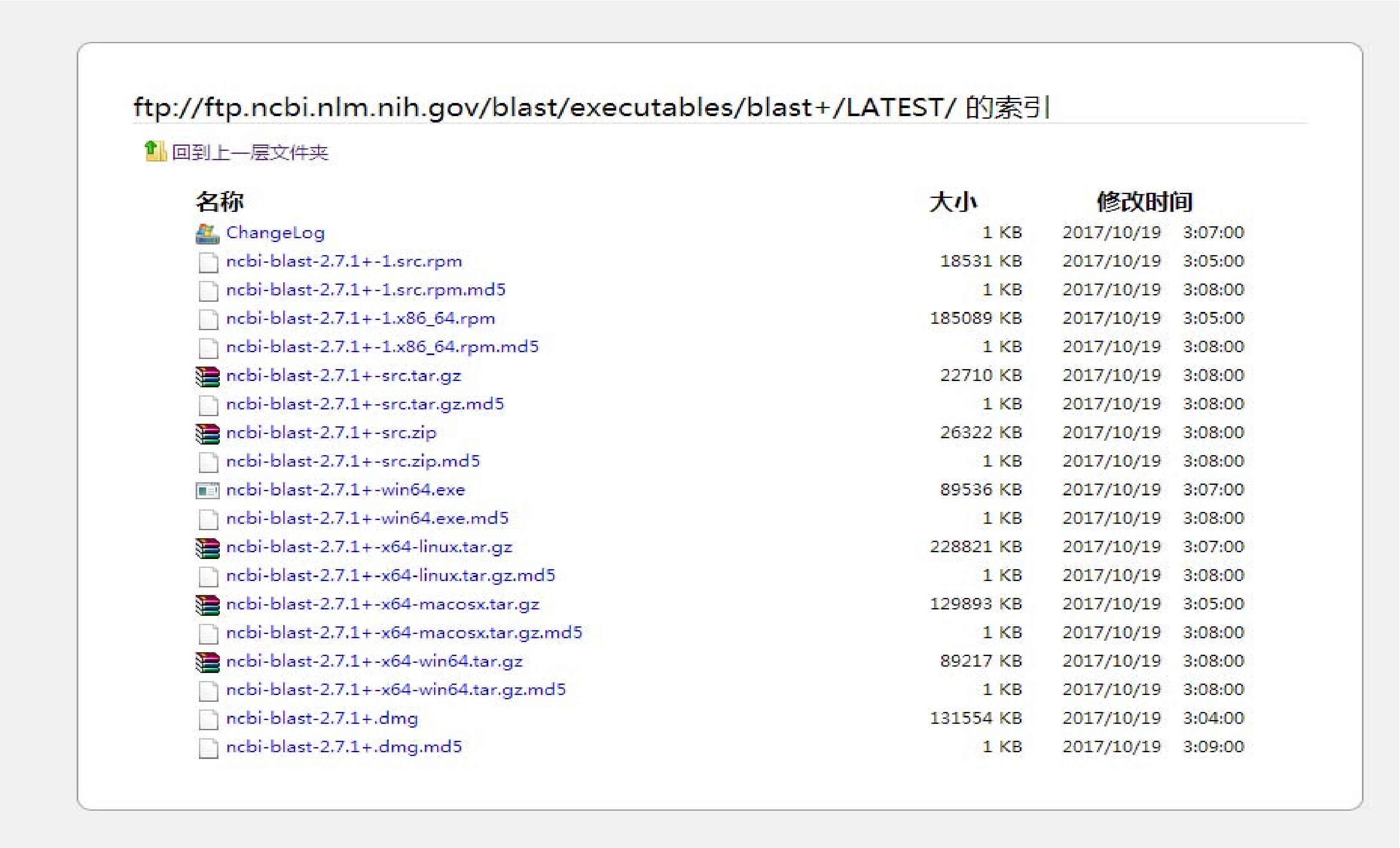
(2) Utilize the standalone BLAST binaries.

You can retrieve BLAST execute files from NCBI ftp sites

ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/

BLAST+

ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/



ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ 的索引

回到上一层文件夹

名称	大小	修改时间	
ChangeLog	1 KB	2017/10/19	3:07:00
ncbi-blast-2.7.1+-1.src.rpm	18531 KB	2017/10/19	3:05:00
ncbi-blast-2.7.1+-1.src.rpm.md5	1 KB	2017/10/19	3:08:00
ncbi-blast-2.7.1+-1.x86_64.rpm	185089 KB	2017/10/19	3:05:00
ncbi-blast-2.7.1+-1.x86_64.rpm.md5	1 KB	2017/10/19	3:08:00
ncbi-blast-2.7.1+-src.tar.gz	22710 KB	2017/10/19	3:08:00
ncbi-blast-2.7.1+-src.tar.gz.md5	1 KB	2017/10/19	3:08:00
ncbi-blast-2.7.1+-src.zip	26322 KB	2017/10/19	3:08:00
ncbi-blast-2.7.1+-src.zip.md5	1 KB	2017/10/19	3:08:00
ncbi-blast-2.7.1+-win64.exe	89536 KB	2017/10/19	3:07:00
ncbi-blast-2.7.1+-win64.exe.md5	1 KB	2017/10/19	3:08:00
ncbi-blast-2.7.1+-x64-linux.tar.gz	228821 KB	2017/10/19	3:07:00
ncbi-blast-2.7.1+-x64-linux.tar.gz.md5	1 KB	2017/10/19	3:08:00
ncbi-blast-2.7.1+-x64-macosx.tar.gz	129893 KB	2017/10/19	3:05:00
ncbi-blast-2.7.1+-x64-macosx.tar.gz.md5	1 KB	2017/10/19	3:08:00
ncbi-blast-2.7.1+-x64-win64.tar.gz	89217 KB	2017/10/19	3:08:00
ncbi-blast-2.7.1+-x64-win64.tar.gz.md5	1 KB	2017/10/19	3:08:00
ncbi-blast-2.7.1+.dmg	131554 KB	2017/10/19	3:04:00
ncbi-blast-2.7.1+.dmg.md5	1 KB	2017/10/19	3:09:00

BLAST use – command line means

List of the BLAST directory

(1) bin – binary files for BLAST execute

(2) doc – documents of BLAST

(3) ncbi_package_info – version information of BLAST

BLAST+

BLAST use

(1) Make a formatted database to use

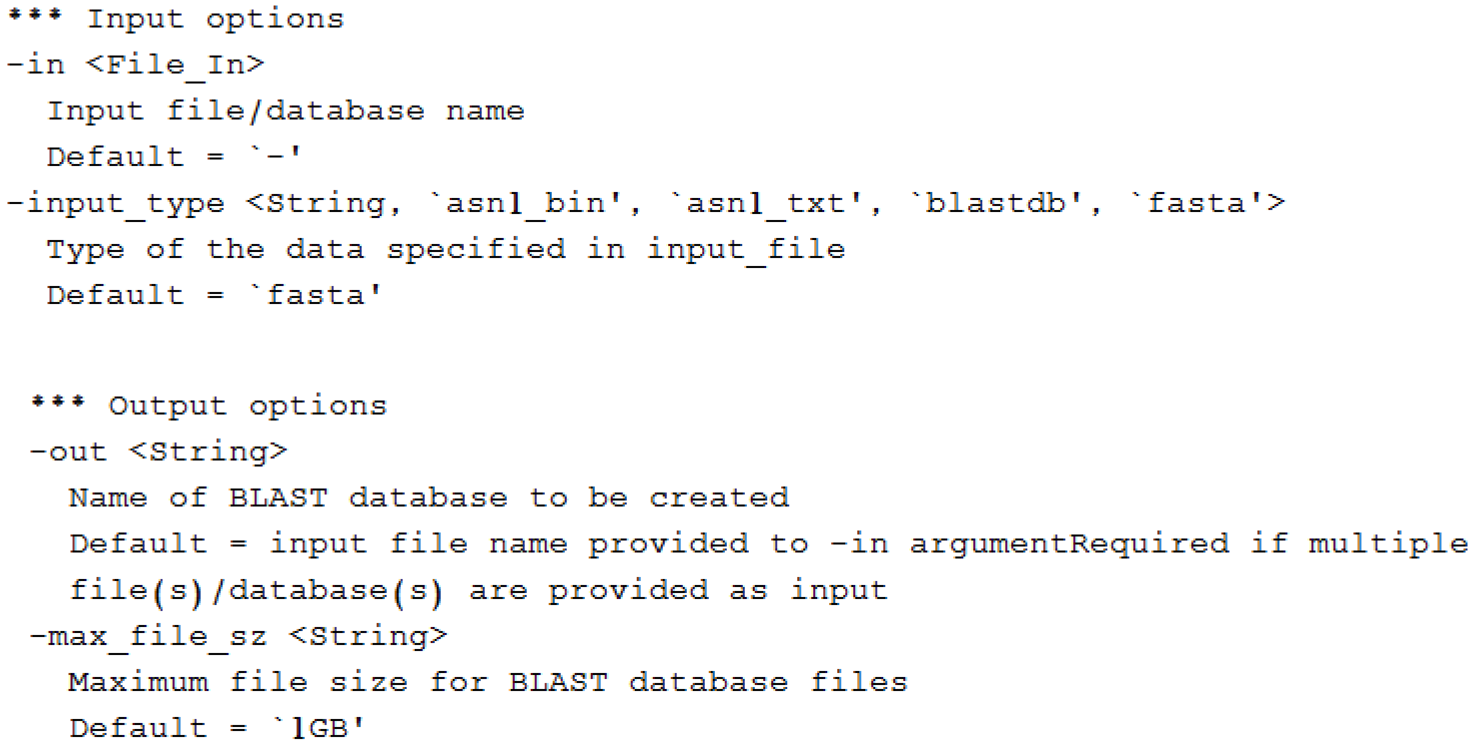
execute command : **makeblastdb**

input: fasta format sequences (database sequences)

output: formatted database , used by BLAST program

BLAST+

Argument of makeblastdb



```
*** Input options
-in <File_In>
  Input file/database name
  Default = `-'
-input_type <String, `asn1_bin', `asn1_txt', `blastdb', `fasta'>
  Type of the data specified in input_file
  Default = `fasta'


 *** Output options
 -out <String>
   Name of BLAST database to be created
   Default = input file name provided to -in argumentRequired if multiple
   file(s)/database(s) are provided as input
 -max_file_sz <String>
   Maximum file size for BLAST database files
   Default = `1GB'
```

BLAST+

BLAST use
(2) Carry out BLAST program

execute command : **blastn, blastp**

input: fasta sequences (query sequences), database
output : query result file

BLAST+

Usage of blastn

```
blastn [-h] [-help] [-import_search_strategy filename]
  [-export_search_strategy filename] [-task task_name] [-db database_name]
  [-dbsize num_letters] [-gilist filename] [-seqidlist filename]
  [-negative_gilist filename] [-entrez_query entrez_query]
  [-db_soft_mask filtering_algorithm] [-db_hard_mask filtering_algorithm]
  [-subject subject_input_file] [-subject_loc range] [-query input_file]
  [-out output_file] [-evalue evalue] [-word_size int_value]
  [-gapopen open_penalty] [-gapextend extend_penalty]
  [-perc_identity float_value] [-xdrop_ungap float_value]
  [-xdrop_gap float_value] [-xdrop_gap_final float_value]
  [-searchsp int_value] [-max_hsps int_value] [-sum_statistics]
  [-penalty penalty] [-reward reward] [-no_greedy]
  [-min_raw_gapped_score int_value] [-template_type type]
  [-template_length int_value] [-dust DUST_options]
  [-filtering_db filtering_database]
  [-window_masker_taxid window_masker_taxid]
  [-window_masker_db window_masker_db] [-soft_masking soft_masking]
  [-ungapped] [-culling_limit int_value] [-best_hit_overhang float_value]
  [-best_hit_score_edge float_value] [-window_size int_value]
  [-off_diagonal_range int_value] [-use_index boolean] [-index_name string]
  [-lcase_masking] [-query_loc range] [-strand strand] [-parse_deflines]
  [-outfmt format] [-show_gis] [-num_descriptions int_value]
  [-num_alignments int_value] [-html] [-max_target_seqs num_sequences]
  [-num_threads int_value] [-remote] [-version]
```

BLAST+

Usage of blastp

```
blastp [-h] [-help] [-import_search_strategy filename]
  [-export_search_strategy filename] [-task task_name] [-db database_name]
  [-dbsize num_letters] [-gilist filename] [-seqidlist filename]
  [-negative_gilist filename] [-entrez_query entrez_query]
  [-db_soft_mask filtering_algorithm] [-db_hard_mask filtering_algorithm]
  [-subject subject_input_file] [-subject_loc range] [-query input_file]
  [-out output_file] [-evalue evalue] [-word_size int_value]
  [-gapopen open_penalty] [-gapextend extend_penalty]
  [-xdrop_ungap float_value] [-xdrop_gap float_value]
  [-xdrop_gap_final float_value] [-searchsp int_value] [-max_hsps int_value]
  [-sum_statistics] [-seg SEG_options] [-soft_masking soft_masking]
  [-matrix matrix_name] [-threshold float_value] [-culling_limit int_value]
  [-best_hit_overhang float_value] [-best_hit_score_edge float_value]
  [-window_size int_value] [-lcase_masking] [-query_loc range]
  [-parse_deflines] [-outfmt format] [-show_gis]
  [-num_descriptions int_value] [-num_alignments int_value] [-html]
  [-max_target_seqs num_sequences] [-num_threads int_value] [-ungapped]
  [-remote] [-comp_based_stats compo] [-use_sw_tback] [-version]
```

THANK YOU